ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THESIS PRESENTED TO
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph. D.

BY
Soroosh REZAZADEH

LOW-COMPLEXITY HIGH PREDICTION ACCURACY VISUAL QUALITY METRICS AND THEIR APPLICATIONS IN H.264/AVC ENCODING MODE DECISION PROCESS

MONTREAL, OCTOBER 9, 2013

BOARD OF EXAMINERS (THESIS PH.D.)

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Stéphane Coulombe, Thesis Supervisor
Department of Software and IT Engineering at École de technologie supérieure

Mr. Mohamed Cheriet, President of the Board of Examiners
Department of Automation Engineering at École de technologie supérieure

Mr. Pierre Dumouchel, Examiner
Department of Software and IT Engineering at École de technologie supérieure

Mr. Guillaume-Alexandre Bilodeau, External Examiner
Department of Computer Engineering at École Polytechnique de Montréal

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

ON SEPTEMBER 6, 2013

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENTS

It would not be possible to complete this PhD thesis without the aid and support of many people over the past years. First and foremost I want to express my gratitude and appreciation towards my supervisor, Prof. Stéphane Coulombe, for his support, thoughtful guidance, useful comments, remarks, and engagement throughout the entire learning process of this PhD thesis. I am grateful to him for giving me the opportunity to become a member of his research group. I also thank the other committee members who have accepted to devote their precious time to evaluating my thesis.

My thanks go also to the members of the jury, Professors Mohamed Cheriet, Pierre Dumouchel, and Guillaume-Alexandre Bilodeau, who have accepted to devote their precious time to evaluating my thesis.

Guillaume-Alexandre Bilodeau, who have accepted to devote their precious time to evaluating my thesis.
I would like to convey my thanks to my friends, including the current and former members of our research group, who helped and provided me a supportive and enjoyable environment in the past years.

I should mention that this research was made possible by the support of Vantrix Corporation and the Natural Sciences and Engineering Research Council of Canada (NSERC) who funded this work under the Collaborative Research and Development Program (NSERC-CRD 326637-05).

Finally, I owe my deepest gratitude to my beloved family, whose boundless love and continuous support and encouragement was the source of inspiration to finish this work.

MÉTRIQUES DE QUALITÉ VISUELLE À HAUTE EXACTITUDE ET À FAIBLE COMPLEXITÉ DE CALCULS ET LEUR APPLICATION AU PROCESSUS DE DÉCISION DE MODES DE L'ENCODEUR H.264/AVC

Soroosh REZAZADEH

RÉSUMÉ

Dans cette thèse, nous développons un nouveau cadre général pour calculer des métriques de qualité d'image avec référence complète dans le domaine des ondelettes discrètes en utilisant l'ondelette de Haar. Le cadre proposé présente un excellent compromis entre l'exactitude et la complexité. Dans notre cadre, les métriques de qualité sont classées soit à base de cartes (map), qui génèrent une carte de qualité (distorsion) dont la contribution à chaque position est mise en commun pour le calcul de la métrique finale, par exemple, la similarité structurelle (SSIM), ou non basées sur des cartes, qui calculent directement la métrique finale, par exemple, le rapport signal sur bruit de crête (PSNR). Pour les métriques basées sur des cartes, le cadre proposé définit une carte de contraste dans le domaine des ondelettes pour la mise en commun des cartes de qualité.

Nous développons aussi une formule permettant de calculer automatiquement le niveau de décomposition en ondelettes approprié pour les métriques basées sur l'erreur en tenant compte de la distance de visualisation désirée. Pour tenir compte de l'effet des détails très fins de l'image dans l'évaluation de la qualité, la méthode proposée définit une carte de contours multi-niveau pour chaque image, qui ne comprend que les sous-bandes d'images les plus informatives.

Pour clarifier l'application du cadre dans le calcul de métriques, nous donnons quelques exemples montrant comment le cadre peut être appliqué pour améliorer la performance de métriques bien connues telles que le SSIM, la fidélité de l'information visuelle (VIF), le PSNR, et la différence absolue. Nous comparons la complexité des différents algorithmes obtenus par le cadre à l'encodage H.264 avec profil de base en utilisant l'implémentation IPP en C/C++ d'Intel. Nous évaluons la performance globale des mesures proposées, y compris leur exactitude de la prédiction, sur deux bases de données de qualité d'image bien connues et une base de données de qualité vidéo. Tous les résultats des simulations confirment l'efficacité du cadre proposé et les mesures d'évaluation de la qualité dans l'amélioration de l'exactitude de la prédiction et aussi la réduction de la complexité de calcul. Par exemple, en utilisant le cadre, nous pouvons calculer le VIF avec environ 5% de la complexité de sa version originale, mais avec une plus grande précision.

Dans la prochaine étape, nous étudions comment le processus de décision de modes de codage en H.264 peut bénéficier des métriques développées. Nous intégrons la métrique SSE_A proposée comme mesure de distorsion dans le processus de décision de mode H.264. Le logiciel de référence H.264/AVC JM est utilisé comme plate-forme de mise en oeuvre et

de validation. Nous proposons un algorithme de recherche pour déterminer la valeur du multiplicateur de Lagrange pour chaque paramètre de quantification (QP). La recherche est appliquée sur trois différents types de séquences vidéo présentant diverses caractéristiques au niveau de l'intensité du mouvement, et les valeurs du multiplicateur de Lagrange qui en résultent sont compilées pour chacun d'eux. Sur la base de notre cadre proposé, nous proposons une nouvelle métrique de qualité $PSNR_A$, et nous l'utilisons dans cette partie (la décision de mode). Les courbes débit-distorsion (RD) simulées montrent que pour le même $PSNR_A$, avec la décision de mode basée SSE_A, le débit est réduit d'environ 5% en moyenne par rapport à l'approche traditionnelle basée SSE sur les séquences avec des niveaux d'intensité de mouvement faibles et moyens. Il est à noter que la complexité de calcul n'est aucunement augmentée en utilisant l'approche basée SSE_A proposée au lieu de la méthode traditionnelle basée SSE. Par conséquent, l'algorithme de décision de mode proposé peut être utilisé pour le codage vidéo en temps réel.

Mots-clés: transformée en ondelettes discrète, évaluation de qualité d'image, système visuel humain (HVS), fidélité de l'information, similarité structurelle, encodage vidéo, H.264, multiplicateur de Lagrange

LOW-COMPLEXITY HIGH PREDICTION ACCURACY VISUAL QUALITY METRICS AND THEIR APPLICATIONS IN H.264/AVC ENCODING MODE DECISION PROCESS

Soroosh REZAZADEH

ABSTRACT

In this thesis, we develop a new general framework for computing full reference image quality scores in the discrete wavelet domain using the Haar wavelet. The proposed framework presents an excellent tradeoff between accuracy and complexity. In our framework, quality metrics are categorized as either map-based, which generate a quality (distortion) map to be pooled for the final score, e.g., structural similarity (SSIM), or non map-based, which only give a final score, e.g., Peak signal-to-noise ratio (PSNR). For map-based metrics, the proposed framework defines a contrast map in the wavelet domain for pooling the quality maps.

We also derive a formula to enable the framework to automatically calculate the appropriate level of wavelet decomposition for error-based metrics at a desired viewing distance. To consider the effect of very fine image details in quality assessment, the proposed method defines a multi-level edge map for each image, which comprises only the most informative image subbands.

To clarify the application of the framework in computing quality scores, we give some examples showing how the framework can be applied to improve well-known metrics such as SSIM, visual information fidelity (VIF), PSNR, and absolute difference. We compare the complexity of various algorithms obtained by the framework to the Intel IPP-based H.264 baseline profile encoding using C/C++ implementations. We evaluate the overall performance of the proposed metrics, including their prediction accuracy, on two well-known image quality databases and one video quality database. All the simulation results confirm the efficiency of the proposed framework and quality assessment metrics in improving the prediction accuracy and also reduction of the computational complexity. For example, by using the framework, we can compute the VIF at about 5% of the complexity of its original version, but with higher accuracy.

In the next step, we study how H.264 coding mode decision can benefit from our developed metrics. We integrate the proposed SSE_A metric as the distortion measure inside the H.264 mode decision process. The H.264/AVC JM reference software is used as the implementation and verification platform. We propose a search algorithm to determine the Lagrange multiplier value for each quantization parameter (*QP*). The search is applied on three different types of video sequences having various motion activity features, and the resulting Lagrange multiplier values are tabulated for each of them. Based on our proposed framework we propose a new quality metric $PSNR_A$, and use it in this part (mode decision). The simulated rate-distortion (RD) curves show that at the same $PSNR_A$, with the SSE_A-based mode decision, the bitrate is reduced about 5% on average compared to the conventional

SSE-based approach for the sequences with low and medium motion activities. It is notable that the computational complexity is not increased at all by using the proposed SSE_A-based approach instead of the conventional SSE-based method. Therefore, the proposed mode decision algorithm can be used in real-time video coding.

Keywords: discrete wavelet transform, image quality assessment, human visual system (HVS), information fidelity, structural similarity, video encoding, H.264, Lagrange multiplier

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

ACR	Absolute Category Rating
AD	Absolute Difference
AVC	Advanced Video Coding
CABAC	Context Adaptive Binary Arithmetic Coding
CALM	Context Adaptive Lagrange Multiplier
CAVLC	Context Adaptive Variable Length Coding
CIF	Common Intermediate Format
CSF	Contrast Sensitivity Function
CW-SSIM	Complex-Wavelet Structural SIMilarity
dB	Decibel
DCT	Discrete Cosine Transform
DPB	Decoded Picture Buffer
DWT	Discrete Wavelet Transform
FF	Fast Fading
FR	Full-Reference
GBlur	Gaussian Blurring
GSM	Gaussian Scale Mixture
GWN	Gaussian White Noise
H.264	Digital video compression and encoding standard
HR	High Rate
HVS	Human Visual System
IFC	Information Fidelity Criterion

IPP	Integrated Performance Primitives
IQM	Image Quality Metrics
JM	Joint Model (H.264 reference software)
JVT	Joint Video Team
KRCC	Kendall Rank Correlation Coefficient
LCC	Linear Correlation Coefficient
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MSE	Mean Squared Errors
MSSIM	Mean Structural SIMilarity
MV	Motion Vector
NR	No-Reference
PSNR	Peak Signal-to-Noise Ratio
QP	Quantization Parameter
RF	Random Field
RD	Rate-Distortion
RDO	Rate-Distortion Optimization
RMSE	Root Mean Square Error
RR	Reduced-Reference
SRCC	Spearman Rank Correlation Coefficient
SAD	Sum of Absolute Differences
SATD	Sum of Absolute Transformed Differences

SSD	Sum of Squared Differences
SSE	Sum of Squared Errors
SSIM	Structural SIMilarity
VDP	Visible Difference Predictor
VIF	Visual Information Fidelity
wSNR	weighted SNR

INTRODUCTION

Problem statement

Image/video coding and transmission systems may introduce some amount of distortion or artifacts in the original/reference signal. The commercial success of an image/video systems depends on their ability to deliver the users good image/video quality consistently. Therefore, image quality assessment plays an important role in the development and validation of various image and video applications, such as compression and enhancement.

The visual quality of an image (or video) is best assessed subjectively by human viewers. But the assessment of subjective quality is time consuming, expensive, and cannot be performed for real-time systems. Therefore, it is essential to define an objective criterion that can measure the difference between the original and the processed image/video signals. Ideally, such an objective measure should correlate well with the perceived difference between two image/video signals, and also varies linearly with the subjective quality.

Many image/video processing systems use mean squared errors (MSE), or equivalently peak signal-to-noise ratio (PSNR), as the objective quality assessment metric due to its simplicity. Because of the non-linear behavior of the human visual system, the PSNR values do not reflect accurately the perceived quality. Therefore, different objective models have arised for accurate visual quality assessment. An overview of the existing quality evaluation models reveals that the computational complexity of assessment techniques that accurately predict quality scores is very high. Owing to the high computational complexity of accurate methods, the PSNR (or MSE) is still used in many image/video processing applications. If we develop low-complexity quality metrics with high-accuracy, it is possible to use them in various real-time image and video processing tasks such as quality control and validation, and video compression with higher perceived quality.

Since the ultimate video quality is judged by human viewers, a well-designed video encoder should be ideally optimized in terms of human visual system perception. Generally speaking, the improvements of visual quality in the video encoding process mainly depend on the following two factors:

- The accuracy of the objective quality assessment models for video sequences.
- The approach to incorporate the objective models into the video encoding framework.

The first factor is addressed in the first part of our thesis, i.e. development of efficient visual quality metrics. Perceptual quality metrics can be adopted and integrated inside several different modules of a video encoder such as mode decision, motion estimation, quantization, and rate control and bit allocation (Su *et al.*, 2012), (Ou, Huang and Chen, 2011), (Yu *et al.*, 2005), (Yuan *et al.*, 2006), (Hrarti *et al.*, 2010).

The macroblock mode decision process is one of the most computationally intensive phases of the video encoding, which also contains within itself the motion estimation during inter-frame predictions. Therefore, it can directly affect the perceptual video quality. Due to its importance, we focus on the mode decision as a potential application of the visual quality metrics. In conventional video encoding mode decision process, sum of squared errors (SSE) is the commonly used metric for measuring the distortion for the ease of calculation. By developing low-complexity visual metrics, the SSE can be substituted with a more accurate metric that is well-correlated with the human evaluations in order to improve the video compression efficiency.

Research objectives

Based upon the shortcomings of the existing methods and the motivations stated in the previous section, we define the main goals of this thesis as follows:

- Develop a general low-complexity framework for full reference visual quality assessment of images (or video frames). This framework must include features for the computation of quality scores not only with lower computational complexity, but also with higher prediction accuracy.
- Create new computational methods for the advanced quality metrics such as SSIM (Wang *et al.*, 2004) and VIF (Sheikh and Bovik, 2006) using the proposed framework. The new models must clearly inherit the mentioned features from the framework, i.e. lower computational complexity and higher prediction accuracy, compared to the original measures.
- Create error-based visual quality metrics for image/video quality assessment using the proposed framework.
- Validate the benefits of the developed quality assessment techniques against state-of-the-art methods by evaluating them on the different image and video quality databases.
- Incorporate the developed quality/distortion metrics, instead of the SSE, in the H.264/AVC mode decision process in order to optimize the video frames' perceptual qualities in terms of the corresponding developed metric.
- Devise an appropriate approach to determine the new Lagrange multiplier values at each *QP* for the incorporated mode decision distortion metric.
- Implement our developed quality/distortion metric in the mode decision process of an H.264/AVC software, and obtain the rate-distortion curves to measure the performance improvement percentage by our proposed mode decision algorithm over the conventional SSE-based approach.

Organization of the thesis

This thesis consists of six main chapters and a conclusion chapter, which all follow this introduction chapter. In the chapter 1, we first introduce the concept of image/video quality assessment, and explain the need for the existence of objective visual quality evaluation models. Then, we explain two major categories in the visual quality evaluation. The

principles and important methods of each category are reviewed briefly, however the new category of top-down approaches are studied in more details due to its importance and also its application in the next chapters. At last, Chapter 1 ends with a discussion section which compares the most important methods against each other and gives the benefits of each of them.

In chapters 2 and 3, we present our proposed framework for calculating the visual quality scores with high prediction accuracy and low computational complexity. In the chapter 2, we describe the principles and theory of our framework, and in the chapter 3, we bring the simulation results of the framework using different image and video databases. Chapter 2 begins with clarifying the shortcomings of the existing quality assessment models and describing the motivations to find a solution approach to improve the performance of these models. After that, each part of our proposed wavelet domain framework is described in details. To show the practical usage of our framework to generate a new quality metric or improvement over an existing model, four different examples are brought afterwards. The first two examples are related to the methods in the category of top-down approaches, i.e. structural similarity approach and the information-theoretic approach. The next two examples are given on the traditional error-based quality models. The first error-based example is using the mean squared error of signals to propose a metric similar to PSNR but with much higher prediction accuracy. The second error-based example uses the absolute differences of the input signals as the distortion measure and benefits from the features of the framework to propose a totally new visual quality metric. It is notable that each of the four given examples is a new and independent visual quality metric on its own.

Chapter 3 shows the simulation results of our framework and the metrics explained in the chapter 2. At first, the computational complexity of different algorithms, working based on the framework, is discussed theoretically and numerically. Then, the prediction accuracy of our metrics is verified using different statistical performance measures. Our tests are performed on three well-known databases: two different image quality databases and a video

quality database. This chapter will finally present concluding remarks on the visual quality metrics based on the simulated results.

In the chapter 4, we review different approaches for macroblock mode decision in video encoding. Inter prediction and various macroblock partitions in H.264/AVC are introduced in the beginning. Then, the Lagrange multiplier estimation is explained as a solution for performing rate distortion optimized mode decision in H.264. After describing the conventional process of encoding a macroblock, important adaptive methods are reviewed to estimate the Lagrange multiplier per video frame. Finally, we study different HVS-based techniques for the Lagrange multiplier calculation. This review provides insights for efficiently using our proposed metrics for mode decision.

Chapter 5 describes our perceptual-based approach for rate-distortion optimized (RDO) mode decision in H.264 Baseline coding. This chapter starts by explaining the pros and cons of existing methods and our motivations behind proposing a new technique for mode decision. Then, the details are given on the proposed approach for the perceptual mode selection. In order to find the best way of determining the corresponding Lagrange multiplier in our method, we analyze the relationships of macroblock distortions in different domains theoretically and empirically. In the next step, we present our proposed search method for the determination of the Lagrange multiplier at each *QP*. After determining the Lagrange multipliers, the simulated rate-distortion (RD) curves are shown for different types of sequences. The RD curves are simulated for both cases of applying adapted and non-adapted Lagrange multipliers. Finally in the concluding section, the potential gains from our improved method are discussed, along with some suggestions for the future research.

In the chapter 6, we list the contributions of our research on the visual quality metrics and H.264 coding mode decision. The last chapter, the conclusion, summarizes the important research results in our thesis and offers the final concluding remarks.

CHAPTER 1

DIGITAL IMAGE AND VIDEO QUALITY ASSESSMENT MODELS AND METRICS

In this chapter, we first explain the general concept of image/video quality assessment. Then, we classify objective visual quality measures, and introduce the main methods of image/video quality evaluation in a nutshell. In the next chapter, we give details of our proposed framework for accurate quality assessment.

1.1 Introduction

As mentioned in the problem statement, an objective criterion is required to measure the level of artifacts and predict the perceptual quality in an image/video processing systems. Objective quality models are usually classified based on the availability of the original image/video signal, which is considered to be of high quality (generally not processed). Generally, quality assessment methods can be classified as full reference (FR) methods, reduced-reference (RR) methods, and no-reference (NR) methods (Wang and Bovik, 2006), as shown in figure 1.1.

FR metrics usually compute the visual quality by comparing every pixel/sample in the distorted image to its corresponding pixel/sample in the original image signal. NR metrics assess the quality of a distorted signal without any reference to the original signal. The NR metrics are usually designed to be application-specific, that is, they directly measure the types of artifacts created by the specific image distortion processes, e.g., blocking by block-based compression and ringing by wavelet-based compression. RR metrics extract some features of both original and distorted signals, and then compare them to give a quality score. They are used when the whole original image/video signal is not available, e.g. in a transmission with a limited bandwidth. In this thesis, we specifically focus on studying and developing the FR quality metrics because they currently lead to higher accuracy.

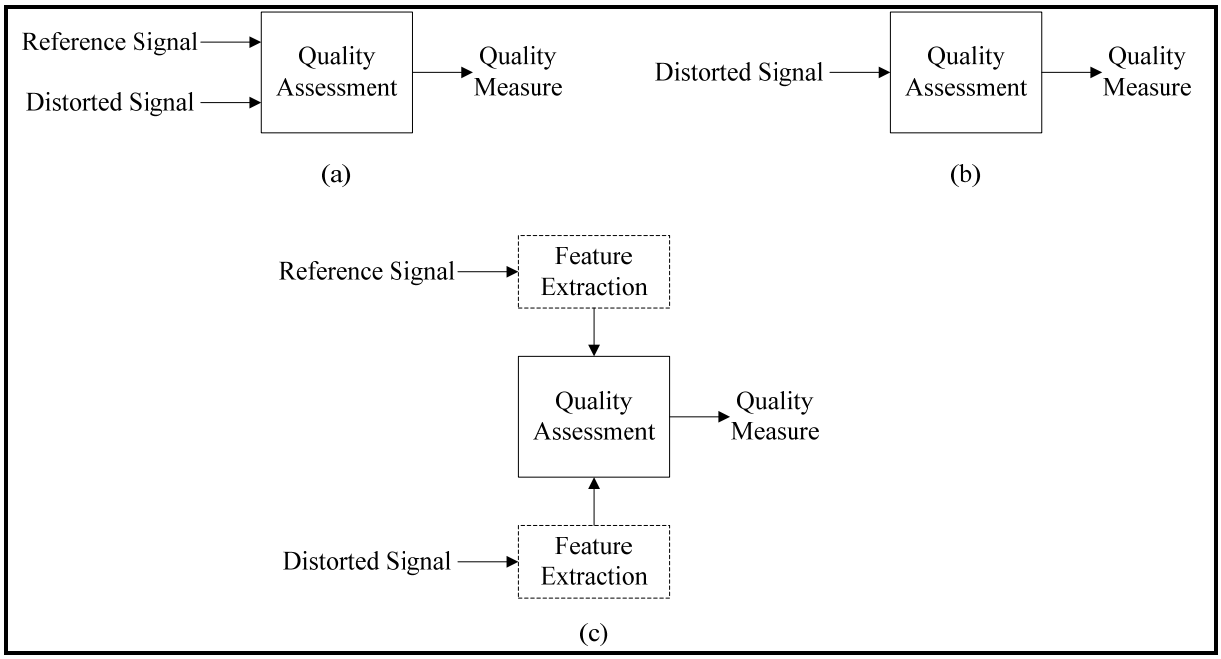

Figure 1.1 Quality assessment approaches: (a) FR method, (b) NR method, (c) RR method.

The mean squared error (MSE) is the most traditional way of measuring the signal fidelity. The goal of a signal fidelity measure is to compare two signals by providing a quantitative score that describes the degree of similarity/ fidelity or, conversely, the level of error/distortion between them (Wang and Bovik, 2009). If we suppose that $\mathbf{X}=\{x_i|i=1,2,\ldots,N\}$ and $\mathbf{Y}=\{y_i|i=1,2,\ldots,N\}$ are two finite-length, discrete signals (e.g., visual images), the MSE between the two signals is defined as in Eq. (1.1).

$$\mathrm{MSE}(\mathbf{X},\mathbf{Y})=\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2 \tag{1.1}$$

where N is the number of signal samples (pixels, if the signals are images), and x_i and y_i are the values of the i^{th} samples in $\mathbf{X}$ and $\mathbf{Y}$ respectively. If one of the signals is considered as an original signal (of acceptable quality), and the other as the distorted version of it whose quality is being evaluated, then the MSE can also be regarded as a measure of signal quality.

In the literature of image/video processing, MSE is usually converted into a peak signal-to-noise ratio (PSNR) measure. The PSNR is more useful than the MSE for comparing images having different dynamic ranges (Wang and Bovik, 2009). The MSE (or equivalently PSNR) is the most widely used objective quality metric in image/video applications. The reason is that the MSE is simple and has a clear physical meaning. It is naturally representative of the error signal energy. In addition, the MSE is an excellent objective metric in the context of optimization and statistical estimation framework. In spite of many favorable properties of MSE, it does not correlate well with the perceived visual quality due to non-linear behavior of the human visual system (HVS).

In (Huynh-Thu and Ghanbari, 2008), the authors have set experiments to investigate where PSNR can or cannot be used as a reliable quality metric. They encoded input sequences at various bit-rates (24-800 kbit/s) with H.264 coding format, and then the decoded sequences were assessed subjectively using ACR international standard method (ITU-T Recommendation P.910). They have found that for a specified content (sequence), the PSNR always monotonically increases with subjective quality as the bit rate increases. Therefore, the PSNR can be used as a performance metric for codec optimization as it correlates highly with subjective quality when the content is fixed. In spite of the existence of monotonic relationship between the PSNR and subjective quality separately per content, it does not exist anymore across different contents. This means different video contents with the same PSNR may have in fact a very different perceptual quality. The PSNR is therefore unreliable as an objective metric for predicting subjective quality. Briefly, the PSNR is not a reliable measure of quality across various video contents, but it is reliable within the content itself (Huynh-Thu and Ghanbari, 2008). Due to deficiencies of PSNR in providing accurate quality predictions, other quality assessment models have been developed by researchers.

Generally speaking, the full reference quality assessment of image and video signals involves two categories of approach: bottom-up and top-down (Wang and Bovik, 2006). In the next sections, we give a brief overview of each approach and introduce main methods in either of them.

1.2 Bottom-up approaches for visual quality assessment

In the bottom-up approaches, perceptual quality scores are best estimated by quantifying the visibility of errors. In order to quantize errors according to the HVS features, techniques in this category try to model the functional properties of different stages of the HVS as characterized by both psychophysical and physiological experiments. This is usually accomplished in several stages of preprocessing, frequency analysis, contrast sensitivity, luminance masking, contrast masking, and error pooling (Wang and Bovik, 2006), (Bovik, 2009), as shown in figure 1.2.

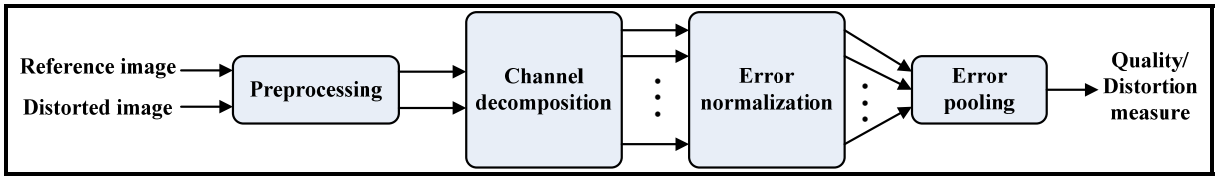

Figure 1.2 A prototypical image quality assessment system based on error visibility. (Adapted from (Wang and Bovik, 2006))

Most of HVS-based quality assessment techniques are multi-channel models, in which each band of spatial frequencies is dealt with by an independent channel. Several important methods of this category have been briefly explained in the annex I.

The bottom-up approaches have several important limitations, which are discussed in (Wang and Bovik, 2006), (Wang *et al.*, 2004). In particular, the HVS is a complex and highly nonlinear system, but most models of early vision are based on linear or quasi-linear operators that have been characterized using restricted and simplistic stimuli. The limited numbers of simple-stimulus experiments are not enough to build a prediction model for perceptual quality that has complex structures. Furthermore, prior information regarding the image content, or attention and fixation likely affect the evaluation of visual quality. These effects are not understood, and are usually ignored by image quality models.

1.3 Top-down approaches for visual quality assessment

The second category of full reference quality assessment includes top-down approaches. In the top-down techniques, the overall functionality of the HVS is considered as a black box, and the input/output relationship is of interest. Thus, top-down approaches do not require any calibration parameters from the HVS or viewing configuration. Two main strategies applied in this category are the structural similarity approach and the information-theoretic approach.

1.3.1 Structural similarity approach

The principal idea underlying the structural similarity approach is that the HVS is highly adapted to extract structural information from visual scenes, and therefore, a measurement of structural similarity (or distortion) should provide a good approximation of the perceptual image quality. In fact, this philosophy considers image degradations as perceived changes in structural information variation. In contrast to error sensitivity concept which is a bottom-up approach and simulating early-stage components in the HVS, structural similarity paradigm is a top-down approach that is mimicking the hypothesized functionality of the overall HVS.

Perhaps the most important method of the structural approach is the Structural SIMilarity (SSIM) index (Wang *et al.*, 2004), which gives an accurate score with acceptable computational complexity compared with other quality metrics (Sheikh, Sabir and Bovik, 2006). SSIM has attracted a great deal of attention in recent years, and has been considered for a wide range of applications. The SSIM is a space domain implementation of the structural similarity idea. In this method, each image can be represented as a vector whose entries are the gray scales of the pixels in the image. The SSIM separates the task of similarity measurement into three comparisons: luminance, contrast and structure. As we know, the luminance of the surface of an object that is imaged or observed is the product of the illumination and the reflectance. The major impacts of illumination changes in an image are variations in the average local luminance and contrast values, but the structures of the objects in the scene are independent of the illumination. Consequently, it is desirable to

separate measurements of luminance and contrast distortions from the other structural distortions that may afflict the image. The diagram of SSIM quality assessment is shown in figure 1.3.

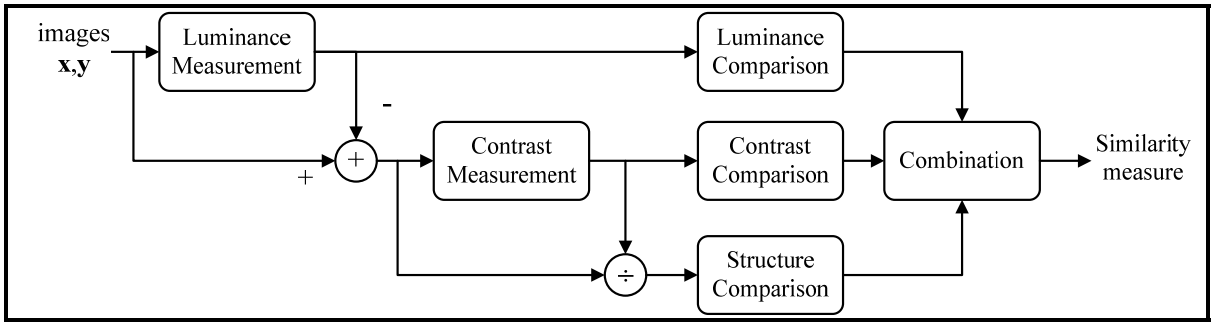

Figure 1.3 Diagram of the SSIM measurement system.
(Adapted from (Wang *et al.*, 2004))

Suppose that **x** and **y** are local image patches taken from the same location of two images that are being compared. As mentioned, the local SSIM index measures the similarities of three elements of the image patches: the luminance similarity $l(\mathbf{x},\mathbf{y})$ of the local patch (brightness values), the contrast similarity $c(\mathbf{x},\mathbf{y})$ of the local patch, and the structural similarity $s(\mathbf{x},\mathbf{y})$ of the local patch. These local similarities are expressed using simple, easily computed statistics, and combined together to form local SSIM:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = l(\mathbf{x},\mathbf{y})\cdot c(\mathbf{x},\mathbf{y})\cdot s(\mathbf{x},\mathbf{y}) = \left(\frac{2\mu_x\mu_y + c_1}{\mu_x^2+\mu_y^2+c_1}\right)\cdot\left(\frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2+\sigma_y^2+c_2}\right)\cdot\left(\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}\right) \quad (1.2)$$

where μ_x and μ_y are (respectively) the local sample means of **x** and **y**, σ_x and σ_y are the local sample standard deviations of **x** and **y**, and σ_{xy} is the sample cross correlation of **x** and **y** after removing their means. The items c_1, c_2, and c_3 are small positive constants that stabilize each term, so that near-zero sample means, variances, or correlations do not lead to numerical instability. In (Wang *et al.*, 2004), robust quality assessment results are obtained using Eq. (1.2) by setting c_3 to $c_2/2$. The SSIM index satisfies the following properties:

- Symmetry: $\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \mathrm{SSIM}(\mathbf{y},\mathbf{x})$

- Boundedness: SSIM($\mathbf{x},\mathbf{y}$) ≤ 1
- Unique maximum: SSIM($\mathbf{x},\mathbf{y}$) = 1 if and only if $\mathbf{x} = \mathbf{y}$

The SSIM index is computed locally (on image blocks or patches). The main reason is that image statistical features are usually highly spatially non-stationary. The local statistics μ_x, σ_x and σ_{xy} as well as the SSIM index, are computed within a local window that moves pixel-by-pixel from the top-left to the bottom-right corner of the image. In (Wang *et al.*, 2004), an 11×11 circular-symmetric Gaussian weighting function (normalized to unit sum) with standard deviation of 1.5 pixels is employed. This choice of window prevents exhibition of blocking artifacts in the SSIM index map and the quality maps show a locally isotropic property. The overall SSIM score of the entire image is then computed by simply averaging the SSIM map:

$$\mathrm{MSSIM}(\mathbf{X},\mathbf{Y}) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{SSIM}(\mathbf{x_j},\mathbf{y_j}) \tag{1.3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the reference and the distorted images, respectively; $\mathbf{x_j}$ and $\mathbf{y_j}$ are the image contents at the jth local window; and M is the number of local windows of the image. The MSSIM is only calculated based on the luminance component of images.

There have been attempts made to improve the SSIM index assessment accuracy. Multi-scale SSIM method (Wang, Simoncelli and Bovik, 2003) for quality assessment provides more flexibility than single-scale SSIM index in incorporating image details at several different resolutions. In this method, an image synthesis-based approach is used to calibrate the parameters that weight the relative importance between different scales. The system diagram for multi-scale SSIM is shown in figure 1.4. The multi-scale SSIM iteratively applies low pass filtering and downsampling up to five different resolutions. The original image is indexed as Scale 1, and the highest scale as Scale M (i.e. 5), which is obtained after (M−1) iterations. At the j^{th} scale, only the contrast comparison $c_j(\mathbf{x},\mathbf{y})$ and the structure comparison $s_j(\mathbf{x},\mathbf{y})$ are calculated. The luminance comparison is computed just at Scale M and is

denoted as $l_{\mathrm{M}}(\mathbf{x},\mathbf{y})$. The overall SSIM evaluation is obtained by combining the measurement at different scales using:

$$\mathrm{SSIM}_{\text{multi-scale}}(\mathbf{x},\mathbf{y}) = [l_{\mathrm{M}}(\mathbf{x},\mathbf{y})]^{\alpha_{\mathrm{M}}} \cdot \prod_{j=1}^{\mathrm{M}} [c_j(\mathbf{x},\mathbf{y})]^{\beta_j} [s_j(\mathbf{x},\mathbf{y})]^{\lambda_j} \tag{1.4}$$

The exponents α_{M}, β_j, and λ_j are used to adjust the relative importance of different components. To simplify parameter selection, we let $\alpha_j = \beta_j = \lambda_j$ for all j's. In addition, the cross-scale settings are normalized such that $\sum_{j=1}^{\mathrm{M}} \lambda_j = 1$. The resulting parameters obtained in (Wang, Simoncelli and Bovik, 2003) according to experiments are: $\beta_1 = \lambda_1 = 0.0448$, $\beta_2 = \lambda_2 = 0.2856$, $\beta_3 = \lambda_3 = 0.3001$, $\beta_4 = \lambda_4 = 0.2363$, and $\alpha_5 = \beta_5 = \lambda_5 = 0.1333$.

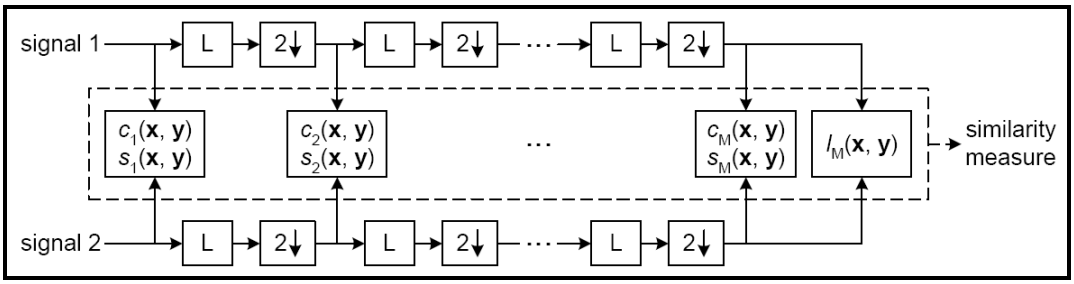

Figure 1.4 Multiscale structural similarity measurement. 2↓ denotes downsampling by 2. (Adapted from (Wang, Simoncelli and Bovik, 2003))

In (Rouse and Hemami, 2008), the authors investigate ways to simplify SSIM in the pixel domain. They study the contribution of each SSIM component in evaluation of common image artifacts. It is finally concluded that by ignoring the mean component in Eq. (1.2) and setting the local average patch values to 128, the linear correlation coefficient is decreased just 1% from the complete computation of the SSIM. The authors in (Yang, Gao and Po, 2008) propose to compute SSIM using subbands at different levels in the discrete wavelet domain. Five-level decomposition using the Daubechies 9/7 wavelet is applied to both the original and the distorted images, and then SSIMs are computed between corresponding subbands. Finally, the similarity score is obtained by the weighted sum of all mean SSIMs.

To determine the weights, a large number of experiments have been performed to measure the sensitivity of the human eye to different frequency bands.

CW-SSIM, which is presented in (Wang and Simoncelli, 2005) and (Sampat *et al.*, 2009), benefits from a complex version of 6-scale, 16-orientation steerable pyramid decomposition characteristics to propose a metric resistant to small geometrical distortions. CW-SSIM is simultaneously insensitive to luminance change, contrast change and spatial translation. The CW-SSIM does not rely on any registration or intensity normalization pre-processing, and just exploits the fact that small translations, scalings, and rotations lead to consistent, describable phase changes in the complex wavelet domain. Yet this method works only when the amount of translation, scaling and rotation is small (compared to the wavelet filter size). In fact, the CW-SSIM is equivalent to applying the spatial domain SSIM index to the magnitudes of the coefficients, where the luminance comparison part is not included since the coefficients are zero-mean (due to the bandpass nature of the wavelet filters). Briefly, the CW-SSIM metric works upon two facts. First, the structural information of local image features is mainly contained in the relative phase patterns of the wavelet coefficients. Second, consistent phase shift of all coefficients does not change the structure of the local image features.

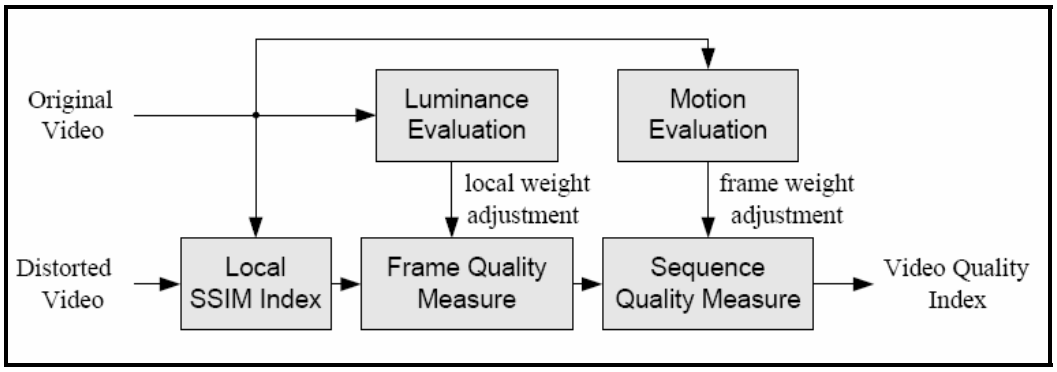

Figure 1.5 SSIM-based video quality assessment system.
(Adapted from (Wang, Lu and Bovik, 2004))

For video quality assessment using structural similarity, SSIM Index is employed as a single measure for various types of distortions. The quality of the distorted video is measured in

three levels as shown in figure 1.5: the local region level, the frame level, and the sequence level (Wang, Lu and Bovik, 2004). Further details about the quality score computation procedure can be found in (Wang, Lu and Bovik, 2004).

For our video quality assessment in next chapters, the image quality metric is applied frame-by-frame on the luminance component of the video, and the overall video quality index is computed as the average of the frame level quality scores. It is known that compression systems, like H.264, produce fairly uniform distortions (or quality) in the video, both spatially and temporally (Seshadrinathan *et al.*, 2010a). Therefore, we suppose that averaging the frames quality scores is an acceptable pooling strategy for assessment of H.264 compression performance.

1.3.2 Information-theoretic approach

In the information-theoretic approach, visual quality assessment is viewed as an information fidelity problem. Methods in this category attempt to relate visual quality to the amount of information that is shared between the images being compared. Shared information is quantified using the mutual information that is a statistical measure of information fidelity. An information fidelity criterion (IFC) for image quality measurement is presented in (Sheikh, Bovik and De Veciana, 2005), which works based on natural scene statistics. In the IFC, the image source is modeled using a Gaussian scale mixture (GSM), while the image distortion process is modeled as an error-prone communication channel. As mentioned, the information shared between the images being compared is quantified using the mutual information. Another information-theoretic quality metric is the visual information fidelity (VIF) index (Sheikh and Bovik, 2006). This index follows the same procedure as the IFC, except that, in the VIF, both the image distortion process and the visual perception process are modeled as error-prone communication channels. The VIF index is one of the most accurate image quality metric, especially when dealing with large image databases (Sheikh, Sabir and Bovik, 2006) and (Wang and Li, 2011). The system diagram of VIF Index is illustrated in figure 1.6.

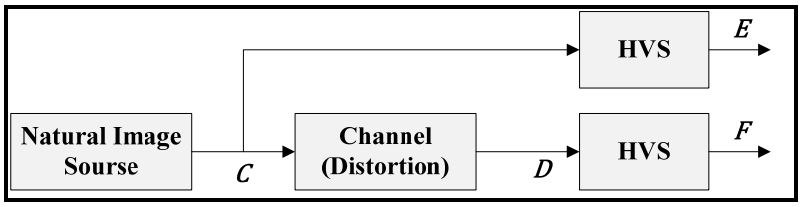

Figure 1.6 VIF index system diagram.
(Adapted from (Sheikh and Bovik, 2006))

In spite of high level of prediction accuracy of VIF, this index has not been given as much consideration as the SSIM index in a variety of applications. This is probably because of its high computational complexity (6.5 times the computation time of the SSIM, index according to (Sheikh and Bovik, 2006)). Most of the complexity in the VIF index comes from over-complete steerable pyramid decomposition, in which the neighboring coefficients from the same subband are linearly correlated. Consequently, the vector Gaussian scale mixture GSM is applied for accurate quality prediction. The decomposed coefficients in each subband are partitioned into non-overlapping blocks of coefficients. Since the blocks do not overlap, each block is assumed to be independent of others, and modeled as a vector. In the next chapter, we propose a low-complexity version of VIF index.

1.4 Discussion

In (Sheikh, Sabir and Bovik, 2006), several prominent full reference image quality measures, from both bottom-up and top-down approaches, were statistically evaluated across a wide variety distortion types. It is reported that VIF index exhibits superior performance relative to all prior methods for image quality assessment, however the performance of the VIF index and the SSIM index are close across a broad diversity of representative image distortion types. As mentioned before, it is more difficult to compute VIF index than SSIM.

Top-down methods are new and fast-evolving compared with bottom-up approaches. Generally, there are some advantages of using top-down methods against bottom-up approaches. Top-down approaches are more accurate and often lead to simpler implantations. Moreover, top-down approaches are more mathematically tractable. For example, the SSIM

index is differentiable, which is useful in gradient-based optimization routines. Both VIF index and SSIM index are bounded between zero and one, while the quality assessment methods belonging to bottom-up approaches do not usually have any specific upper and lower bounds.

CHAPTER 2

THE PROPOSED LOW-COMPLEXITY DISCRETE WAVELET TRANSFORM FRAMEWORK FOR FULL REFERENCE IMAGE QUALITY ASSESSMENT

2.1 Motivation

In this chapter, we propose a novel general framework to calculate image quality metrics (IQM) in the discrete wavelet domain. The proposed framework can be applied to both top-down and error-based (bottom-up) approaches, as explained in subsequent sections. This framework can be applied to map-based metrics, which generate quality (or distortion) maps such as the SSIM map and the absolute difference (AD) map, or to non map-based ones, which give a final score such as the PSNR and the VIF index. We also show that, for these metrics, the framework leads to improved accuracy and reduced complexity compared to the original metrics.

We developed the new framework mainly because of the following shortcomings of the current methods. First, the computational complexity of assessment techniques that accurately predict quality scores is very high. If we develop high-accuracy low complexity quality metrics, it is possible to use them in real-time image/video processing applications such as motion estimation and video encoding rate control (Yasakethu *et al.*, 2008), can be solved more efficiently if an accurate low complexity quality metric is used. Second, the bottom-up approach reviewed (Teo and Heeger, 1994),(Chandler and Hemami, 2007) specifies that those techniques apply the multi-resolution transform, decomposing the input image into large numbers of resolutions (five or more). As the HVS is a complex system that is not completely known to us, combining the various bands into the final metric is difficult. In similar top-down methods, such as multi-scale and multi-level SSIMs (Wang, Simoncelli and Bovik, 2003), (Yang, Gao and Po, 2008), determining the sensitivity of the HVS to different scales or subbands requires many experiments. Our new approach does not require such heavy experimentation to determine parameters. Third, top-down methods, like SSIM, gather local statistics within a square sliding window, and the computed statistics of image

blocks in the wavelet domain are more accurate. Fourth, a large number of decomposition levels, as in (Yang, Gao and Po, 2008), would make the size of the approximation subband very small, so it would no longer be able to help in extracting image statistics effectively. In contrast, the approximation subband contains the main image content, and we have observed that this subband has a major impact on improving quality prediction accuracy. Fifth, previous SSIM methods use the mean of the quality map to give the overall image quality score. However, distortions in various image areas have different impacts on the HVS. In our framework, we introduce a contrast map in the wavelet domain for pooling quality maps.

In the rest of this chapter, we first describe our proposed general framework for image quality assessment. Then, we explain how the proposed framework is used to calculate the currently well-known objective quality metrics.

2.2 The proposed discrete wavelet domain image quality assessment framework

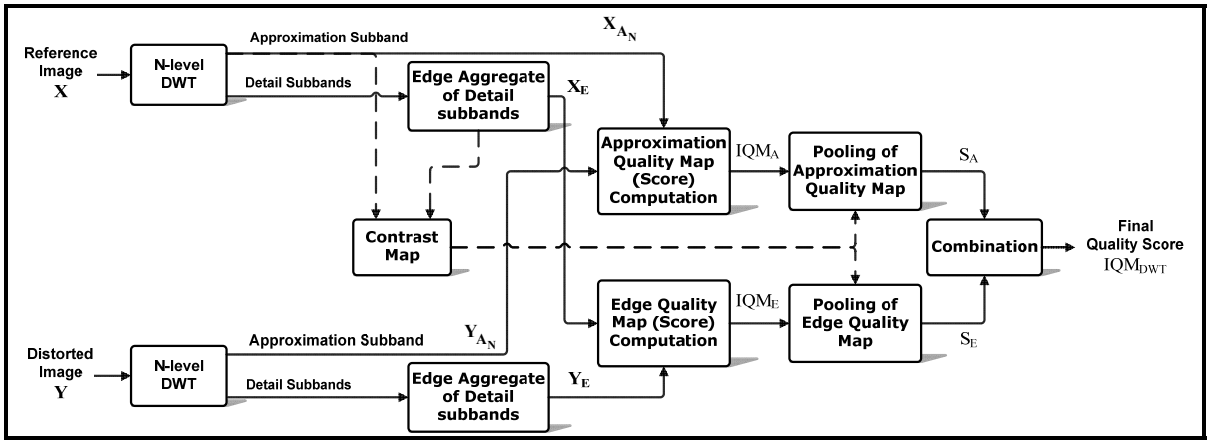

Figure 2.1 Block diagram of the proposed discrete wavelet domain image quality assessment framework.

In this section, we describe a DWT-based framework for computing a general purpose FR image quality metric (IQM). The block diagram of the proposed framework is shown in figure 2.1. The dashed lines in this figure display the parts that may be omitted based on whether or not it is a map-based metric. Let **X** and **Y** denote the reference and distorted

images respectively. The procedure for calculating the proposed version of IQM is set out and explained in the following steps.

Step 1. We perform N-level DWT on both the reference and the distorted images based on the Haar wavelet filter. With N-level decomposition, the approximation subbands $\mathbf{X}_{A_N}$ and $\mathbf{Y}_{A_N}$, as well as a number of detail subbands, are obtained.

The Haar wavelet is the oldest and the simplest wavelet. Haar wavelet has very interesting properties including symmetry, orthogonality, biorthogonality, and compact support (Daubechies, 1992). Since the Haar wavelet is orthogonal, the frequency components of input signal can be analyzed. The length of low-pass and high-pass filters for the Haar wavelet is 2. The Haar decomposition low-pass filter Lo_D and the decomposition high-pass filter Hi_D are defined as Lo_D=$\left(1/\sqrt{2}\right)[1\ 1]$ and Hi_D=$\left(1/\sqrt{2}\right)[-1\ 1]$, respectively (Daubechies, 1992). For discrete 2-D (two dimensional) wavelet decomposition, these filters are applied to rows, and then to columns of the image signal. There is no need for multiplications in Haar transform. It requires only additions and a final scaling, so the computation time is very short. In fact, we just perform a simple averaging of the original signal to obtain the approximation subband. For example, for a single decomposition level, if the pixel intensities in a 2×2 window are assumed to be P_1, P_2, P_3, and P_4, the resulting approximation coefficient would be $0.5(P_1+P_2+P_3+P_4)$. Since a kind of averaging process is performed within the HVS when looking at the visual scenes, the characteristics of the Haar wavelet helps to emulate this feature of HVS for quality assessment.

The Haar wavelet has been used previously in some quality assessment and compression methods (Bolin and Meyer, 1999),(Lai and Kuo, 2000). For our framework, we chose the Haar filter for its simplicity and good performance. The Haar wavelet has very low computational complexity compared to other wavelets. In addition, based on our simulations, it provides more accurate quality scores than other wavelet bases. The reason for this is that symmetric Haar filters have a generalized linear phase, so the perceptual image structures can

be preserved. Also, Haar filters can avoid over-filtering the image, owing to their short filter length.

The number of levels (N) selected for structural or information-theoretic strategies, such as SSIM or VIF, is equal to one. The reason for this is that, for more than one level of decomposition, the resolution of the approximation subband is reduced exponentially and it becomes very small. Consequently, a large number of important image structures or information will be lost in that subband. But, for error-based approaches, like PSNR or absolute difference (AD), we can formulate the required decomposition levels N as follows: when an image is viewed at distance d from a display of height h, we have (Wang, Ostermann and Zhang, 2002):

$$f_\theta = \frac{\pi}{180}\frac{\mathrm{d}}{\mathrm{h}} f_s \quad \text{(cycles/degree)} \tag{2.1}$$

where f_θ is the angular frequency that has a cycle/degree (cpd) unit; and f_s denotes the spatial frequency. For an image of height H, the Nyquist theorem results in Eq. (2.2):

$$\left(f_s\right)_{max} = \frac{\mathrm{H}}{2} \quad \text{(cycles/picture height)} \tag{2.2}$$

It is known that the HVS has a peak response for frequencies at about 2-4 cpd. We chose f_θ = 3. If the image is assessed at a viewing distance of $\mathrm{d} = k\mathrm{h}$, using Eq. (2.1) and Eq. (2.2), we deduce Eq. (2.3):

$$\mathrm{H} \geq \frac{360 f_\theta}{\pi(\mathrm{d/h})} = \frac{360\times 3}{3.14\times k} \approx \frac{344}{k} \tag{2.3}$$

So, the effective size of an image for human eye assessment should be around (344/k). Accordingly, the minimum size of the approximation subband after N-level decomposition should be approximately equal to (344/k). For an image of size H×W, N is calculated as follows (considering that N must be non negative):

$$\frac{\min(\mathrm{H},\mathrm{W})}{2^{\mathrm{N}}} \approx \frac{344}{k} \Rightarrow \mathrm{N} = \mathrm{round}\left(\log_2 \left(\frac{\min(\mathrm{H},\mathrm{W})}{(344/k)} \right) \right) \tag{2.4}$$

$$\mathrm{N} \geq 0 \Rightarrow \mathrm{N} = \max\left(0, \mathrm{round}\left(\log_2 \left(\frac{\min(\mathrm{H},\mathrm{W})}{(344/k)} \right) \right) \right) \tag{2.5}$$

Step 2. We calculate the quality map (or score) by applying IQM between the approximation subbands of $\mathbf{X}_{A_N}$ and $\mathbf{Y}_{A_N}$, and call it the approximation quality map (or score), IQM_A. Examples of IQM computations applied to various quality metrics, such as SSIM and VIF, will be presented in the next section.

Step 3. An estimate of the image edges is formed for each image using an aggregate of detail subbands. If we apply the N-level DWT to the images, the edge map (estimate) of image **X** is defined as:

$$\mathbf{X}_{\mathrm{E}}(m,n) = \sum_{\mathrm{L}=1}^{\mathrm{N}} \mathbf{X}_{\mathrm{E,L}}(m,n) \tag{2.6}$$

where $\mathbf{X}_{\mathrm{E}}$ is the edge map of **X**; and $\mathbf{X}_{\mathrm{E,L}}$ is the image edge map at decomposition level L, computed as defined in Eq. (2.7). In Eq. (2.7), $\mathbf{X}_{\mathrm{H_L}}$, $\mathbf{X}_{\mathrm{V_L}}$, and $\mathbf{X}_{\mathrm{D_L}}$ denote the horizontal, vertical, and diagonal detail subbands obtained at the decomposition level L for image **X** respectively. $\mathbf{X}_{\mathrm{H_L,A_{N-L}}}$, $\mathbf{X}_{\mathrm{V_L,A_{N-L}}}$, and $\mathbf{X}_{\mathrm{D_L,A_{N-L}}}$ are the wavelet packet approximation subbands obtained by applying an (N−L)-level DWT on $\mathbf{X}_{\mathrm{H_L}}$, $\mathbf{X}_{\mathrm{V_L}}$, and $\mathbf{X}_{\mathrm{D_L}}$ respectively. The parameters μ, λ, and ψ are constant. As the HVS is more sensitive to the horizontal and vertical subbands and less sensitive to the diagonal one, greater weight is given to the horizontal and vertical subbands. We arbitrarily propose $\mu = \lambda = 4.5\psi$ in this paper, which results in $\mu = \lambda = 0.45$ and $\psi = 0.10$ to satisfy Eq. (2.8).

$$\mathbf{X}_{\mathrm{E,L}}(m,n)=\begin{cases}\sqrt{\mu\cdot\left(\mathbf{X}_{\mathrm{H_L}}(m,n)\right)^2+\lambda\left(\mathbf{X}_{\mathrm{V_L}}(m,n)\right)^2+\psi\left(\mathbf{X}_{\mathrm{D_L}}(m,n)\right)^2} & \text{if} \quad \mathrm{L=N}\\ \sqrt{\mu\cdot\left(\mathbf{X}_{\mathrm{H_L,A_{N\text{-}L}}}(m,n)\right)^2+\lambda\left(\mathbf{X}_{\mathrm{V_L,A_{N\text{-}L}}}(m,n)\right)^2+\psi\left(\mathbf{X}_{\mathrm{D_L,A_{N\text{-}L}}}(m,n)\right)^2} & \text{if} \quad \mathrm{L<N}\end{cases} \tag{2.7}$$

$$\mu+\lambda+\psi=1 \tag{2.8}$$

$\mathbf{X}_{\mathrm{A_2}}$	$\mathbf{X}_{\mathrm{H_2}}$	$\mathbf{X}_{\mathrm{H_1,A_1}}$	$\mathbf{X}_{\mathrm{H_1,H_1}}$
$\mathbf{X}_{\mathrm{V_2}}$	$\mathbf{X}_{\mathrm{D_2}}$	$\mathbf{X}_{\mathrm{H_1,V_1}}$	$\mathbf{X}_{\mathrm{H_1,D_1}}$
$\mathbf{X}_{\mathrm{V_1,A_1}}$	$\mathbf{X}_{\mathrm{V_1,H_1}}$	$\mathbf{X}_{\mathrm{D_1,A_1}}$	$\mathbf{X}_{\mathrm{D_1,H_1}}$
$\mathbf{X}_{\mathrm{V_1,V_1}}$	$\mathbf{X}_{\mathrm{V_1,D_1}}$	$\mathbf{X}_{\mathrm{D_1,V_1}}$	$\mathbf{X}_{\mathrm{D_1,D_1}}$

Figure 2.2 The wavelet subbands for a two-level decomposed image.

The edge map of **Y** is defined in a similar way for **X**. As an example, Figure 2.2 depicts the subbands of image **X** for N=2. The subbands involved in computing the edge map are shown in color in this figure. It is notable that the edge map is intended to be an estimate of image edges. Thus, the most informative subbands are used in forming the edge map, rather than considering all of them. It is notable that edge maps of different subbands are of the same size, so they can be combined together without any problem.

In our method, we use only 3N edge bands. If we considered all the bands in our edge map, we would have to use $4^{\mathrm{N}}-1$ bands. When N is greater than or equal to 2, the value $4^{\mathrm{N}}-1$ is much greater than 3N. Thus, our proposed edge map helps save computation effort. According to our simulations, considering all the image subbands in calculating the edge map does not have a significant impact on increasing prediction accuracy. It is notable that the edge maps only reflect the fine-edge structures of images.

Step 4. We apply IQM between the edge maps $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{Y}_{\mathrm{E}}$. The resulting quality map (or score) is called the edge quality map (or score), $\mathrm{IQM}_{\mathrm{E}}$.

Step 5. Some metrics, like AD or SSIM, generate an intermediate quality map which should be pooled to reach the final score. In this step, we form a contrast map function for pooling the approximation and edge quality maps. It is well known that the HVS is more sensitive to areas near the edges (Wang and Bovik, 2006). Therefore, the pixels in the quality map near the edges should be given more importance. At the same time, high-energy (or high-variance) image regions are likely to contain more information to attract the HVS (Wang and Shang, 2006). Thus, the pixels of a quality map in high-energy regions must also receive higher weights (more importance). Based on these facts, we can combine our edge map with the computed variance to form a contrast map function. The contrast map is computed within a local Gaussian square window, which moves (pixel by pixel) over the entire edge maps $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{Y}_{\mathrm{E}}$. As in (Wang *et al.*, 2004), we define a Gaussian sliding window $\mathbf{W} = \{w_k \,|\, k = 1, 2, \ldots, \mathrm{K}\}$ with a standard deviation of 1.5 samples, normalized to unit sum. Here, we set the number of coefficients K to 16, that is, a 4×4 window. This window size is not too large and can provide accurate local statistics. The contrast map is defined as follows:

$$Contrast(\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{A_N}}) = (\mu_{x_{\mathrm{E}}}^2 \sigma_{x_{\mathrm{A_N}}}^2)^{0.15} \tag{2.9}$$

$$\sigma_{x_{\mathrm{A_N}}}^2 = \sum_{k=1}^{\mathrm{K}} w_k (x_{\mathrm{A_N},k} - \mu_{x_{\mathrm{A_N}}})^2 \tag{2.10}$$

$$\mu_{x_{\mathrm{E}}} = \sum_{k=1}^{\mathrm{K}} w_k x_{\mathrm{E},k} \quad , \quad \mu_{x_{\mathrm{A_N}}} = \sum_{k=1}^{\mathrm{K}} w_k x_{\mathrm{A_N},k} \tag{2.11}$$

where $\mathbf{x}_{\mathrm{E}}$ and $\mathbf{x}_{\mathrm{A_N}}$ denote image patches of $\mathbf{X}_{\mathrm{E}}$ and $\mathbf{X}_{\mathrm{A_N}}$ within the sliding window. It is notable that the contrast map merely exploits the original image statistics to form the weighted function for quality map pooling and statistics of the distorted image are not used.

The reason is that structures of the image may change due to distortion, so the statistics of distorted image are not very accurate.

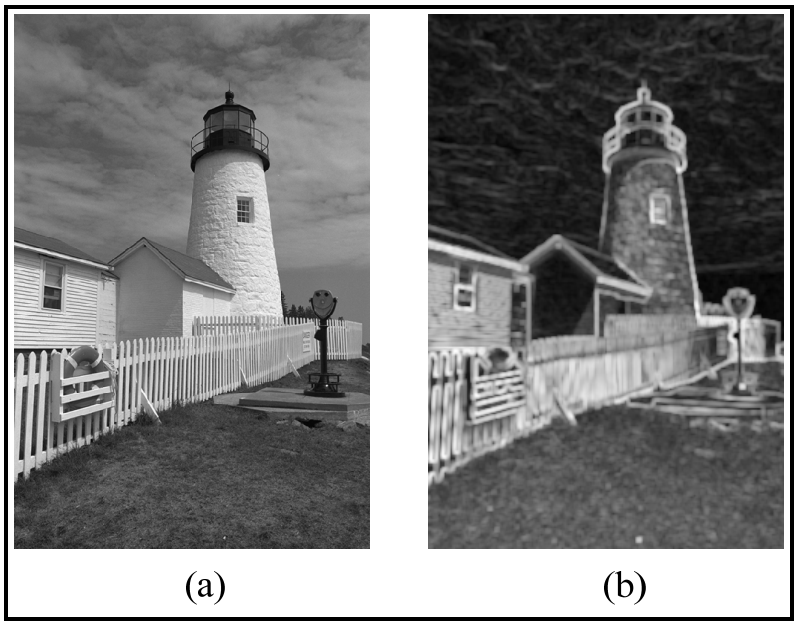

Figure 2.3 (a) Original image; (b) Contrast map computed using Eq. (2.9). The sample values of the contrast map are scaled between [0,255] for easy observation.

Figure 2.3(b) demonstrates the resized contrast map obtained by Eq. (2.9) for a typical image in figure 2.3(a). As can be seen in figure 2.3, the contrast map nicely shows the edges and important image structures to the HVS. Brighter (higher) sample values in the contrast map indicate image structures that are more important to the HVS and play an important role in judging image quality.

Step 6. For map-based metrics, the contrast map in Eq. (2.9) is used for weighted pooling of the approximation quality map $\mathrm{IQM_A}$ and the edge quality map $\mathrm{IQM_E}$.

$$S_A = \frac{\sum_{j=1}^{M} Contrast(\mathbf{x}_{\mathrm{E},j}, \mathbf{x}_{\mathrm{A_N},j}) \cdot \mathrm{IQM_A}(\mathbf{x}_{\mathrm{A_N},j}, \mathbf{y}_{\mathrm{A_N},j})}{\sum_{j=1}^{M} Contrast(\mathbf{x}_{\mathrm{E},j}, \mathbf{x}_{\mathrm{A_N},j})} \tag{2.12}$$

$$S_E = \frac{\sum_{j=1}^{M} Contrast(\mathbf{x}_{E,j}, \mathbf{x}_{A_N,j}) \cdot \mathrm{IQM}_E(\mathbf{x}_{E,j}, \mathbf{y}_{E,j})}{\sum_{j=1}^{M} Contrast(\mathbf{x}_{E,j}, \mathbf{x}_{A_N,j})} \tag{2.13}$$

where $\mathbf{x}_{E,j}$ and $\mathbf{x}_{A_N,j}$ in the contrast map function denote image patches in the j-th local window; $\mathbf{x}_{A_N,j}$, $\mathbf{y}_{A_N,j}$, $\mathbf{x}_{E,j}$, and $\mathbf{y}_{E,j}$ in the quality map (or score) terms are image patches (or pixels) in the j-th local window position; M is the number of samples (pixels) in the quality map; and S_A and S_E represent the approximation and edge quality scores respectively. It is notable that, for non map-based metrics like PSNR, $S_A = \mathrm{IQM_A}$ and $S_E = \mathrm{IQM_E}$.

Step 7. Finally, the approximation and edge quality scores are combined linearly, as defined in Eq. (2.14), to obtain the overall quality score $\mathrm{IQM_{DWT}}$ between images **X** and **Y**:

$$\mathrm{IQM_{DWT}}(\mathbf{X}, \mathbf{Y}) = \beta \cdot S_A + (1-\beta) \cdot S_E \qquad 0 < \beta \le 1 \tag{2.14}$$

where $\mathrm{IQM_{DWT}}$ gives the final quality score between the images; and β is a constant. As the approximation subband contains the main image contents, β should be close to 1 to give the approximation quality score (S_A) much greater importance. We set β to 0.85 in our simulations, which means the approximation quality score constitutes 85% of the final quality score and only 15% is made up of the edge quality score.

2.3 Examples of framework applications

In this section, we clarify how to apply a framework to various well-known quality assessment methods. The SSIM and VIF methods are explained in the top-down category (Rezazadeh and Coulombe, 2009),(Rezazadeh and Coulombe, 2010), and the PSNR and AD approaches are discussed in the error-based (bottom-up) category. The SSIM and AD metrics are examples of map-based metrics, and VIF and PSNR are examples of non map-based metrics.

2.3.1 Structural SIMilarity

In the first step, we need to make sure that one decomposition level, as we previously proposed, works appropriately for calculating the proposed $SSIM_{DWT}$ value. Since the image approximation subband plays the major role in our algorithm, we want to determine N in such a way that it maximizes the prediction accuracy of the approximation quality score $SSIM_A$ by itself. So, using an approximation quality part with computational complexity that is lower than that of the full metric helps to predict quality accurately. The plots in figure 2.4 show the linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SRCC) between $SSIM_A$ and the mean opinion score (MOS) values for different N. In performing this test, all the distorted images of the IVC image database (Le Callet and Autrusseau, 2005) were included in computing the LCC and SRCC. The distorted images in this database were generated from 10 original images using 4 different processing methods: JPEG, JPEG2000, LAR coding, and Blurring (Le Callet and Autrusseau, 2005). As can be seen from figure 2.4, $SSIM_A$ achieves its best performance for N=1.

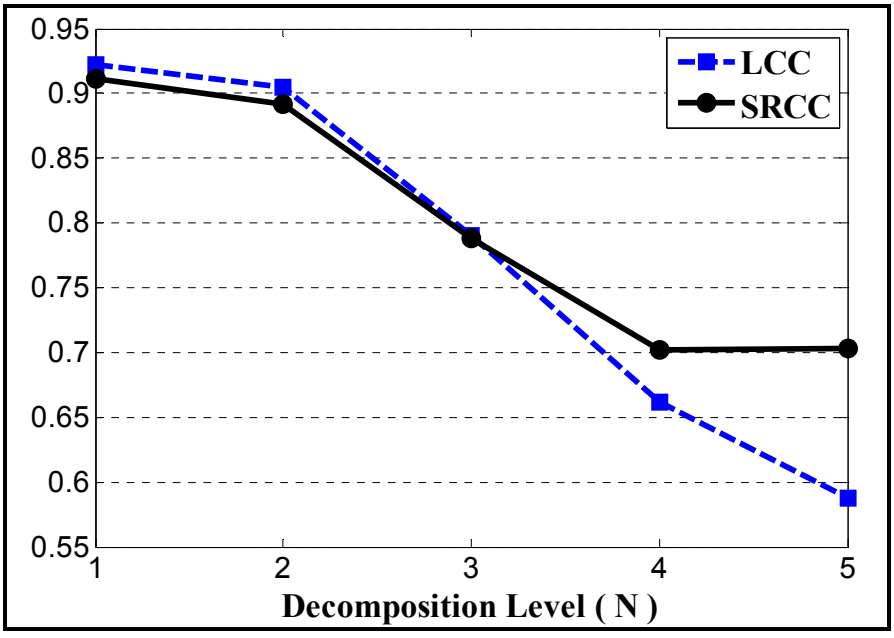

Figure 2.4 LCC and SRCC between the MOS and mean $SSIM_A$ prediction values for various decomposition levels.

In the second step, we calculate the approximation SSIM map, SSIM_A, between the approximation subbands of $\mathbf{X}$ and $\mathbf{Y}$. For each image patch $\mathbf{x}_A$ and $\mathbf{y}_A$ within the first level approximation subbands of $\mathbf{X}$ and $\mathbf{Y}$, SSIM_A is computed as Eq. (2.15):

$$\mathrm{SSIM}_A(\mathbf{x}_A, \mathbf{y}_A) = \mathrm{SSIM}(\mathbf{x}_A, \mathbf{y}_A) \tag{2.15}$$

The SSIM map is calculated according to the method in (Wang *et al.*, 2004). We keep all the parameters the same as those proposed in (Wang *et al.*, 2004), except for window size, for which we use a sliding 4×4 Gaussian window. In the third step, the edge-map function is defined for each image using Eqs. (2.6), (2.7):

$$\mathbf{X}_E(m,n) = \sqrt{0.45 \cdot \left(\mathbf{X}_{H_1}(m,n)\right)^2 + 0.45 \cdot \left(\mathbf{X}_{V_1}(m,n)\right)^2 + 0.1 \cdot \left(\mathbf{X}_{D_1}(m,n)\right)^2} \tag{2.16}$$

$$\mathbf{Y}_E(m,n) = \sqrt{0.45 \cdot \left(\mathbf{Y}_{H_1}(m,n)\right)^2 + 0.45 \cdot \left(\mathbf{Y}_{V_1}(m,n)\right)^2 + 0.1 \cdot \left(\mathbf{Y}_{D_1}(m,n)\right)^2} \tag{2.17}$$

where (m,n) shows the sample position within the wavelet subbands.

In the fourth step, the edge SSIM map, SSIM_E, is calculated between two images using the following formula:

$$\mathrm{SSIM}_E(\mathbf{x}_E, \mathbf{y}_E) = \frac{2\sigma_{x_E, y_E} + c}{\sigma_{x_E}^2 + \sigma_{y_E}^2 + c} \tag{2.18}$$

$$c = (kL)^2 \quad , \qquad k \ll 1 \tag{2.19}$$

where σ_{x_E, y_E} is the covariance between image patches $\mathbf{x}_E$ and $\mathbf{y}_E$ (of $\mathbf{X}_E$ and $\mathbf{Y}_E$); parameters $\sigma_{x_E}^2$ and $\sigma_{y_E}^2$ are variances of $\mathbf{x}_E$ and $\mathbf{y}_E$ respectively; k is a small constant; and L is a dynamic range of pixels (255 for gray-scale images). The correlation coefficient and variances are computed in the same manner as presented in (Wang *et al.*, 2004). In fact, as the edge map

only forms image-edge structures and contains no luminance information, the luminance comparison part of the SSIM map in (Wang *et al.*, 2004) is omitted for the edge SSIM map.

In the next steps, the contrast map is obtained using Eq. (2.9) for pooling $SSIM_A$ and $SSIM_E$ in Eqs. (2.12),(2.13). The final quality score, $SSIM_{DWT}$, is calculated using Eq. (2.14):

$$\mathrm{SSIM}_{\mathrm{DWT}}(\mathbf{X},\mathbf{Y}) = \beta \cdot S_A + (1-\beta) \cdot S_E \tag{2.20}$$

2.3.2 Visual information fidelity

As mentioned in the previous chapter, due to high computational complexity of VIF index, this metric this index has not been given as much consideration as the SSIM index in a variety of applications. In this subsection, we explain the steps for calculating VIF in the discrete wavelet domain by exploiting the proposed framework. The proposed approach is more accurate than the original VIF index, and yet is less complex than the VIF index. It applies real Cartesian-separable wavelets and uses scalar GSM instead of vector GSM in modeling the images for VIF computation.

2.3.2.1 Scalar GSM-based VIF

Scalar GSM has been described and applied in the computation of the IFC (Sheikh, Bovik and De Veciana, 2005). We repeat that procedure here for VIF index calculation using scalar GSM. Considering figure 1.6, let $\mathrm{C}^{\mathrm{M}} = (\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_{\mathrm{M}})$ denote M elements from C, and let $\mathrm{D}^{\mathrm{M}} = (\mathrm{D}_1, \mathrm{D}_2, \ldots, \mathrm{D}_{\mathrm{M}})$ be the corresponding M elements from D. C and D denote the RFs from the reference and distorted signals respectively (as in (Sheikh, Bovik and De Veciana, 2005), the models correspond to one subband). C is a product of two stationary random fields (RFs) that are independent of each other:

$$C = \{\mathrm{C}_i : i \in \mathrm{I}\} = S \cdot U = \{\mathrm{S}_i \cdot \mathrm{U}_i : i \in \mathrm{I}\} \tag{2.21}$$

where I denotes the set of spatial indices for the random field (RF); S is an RF of positive scalars; and U is a Gaussian scalar RF with zero mean and variance σ_U^2. The distortion model is a signal attenuation and additive Gaussian noise, defined as follows:

$$D = \{\mathrm{D}_i : i \in \mathrm{I}\} = GC + V = \{g_i \mathrm{C}_i + \mathrm{V}_i : i \in \mathrm{I}\} \tag{2.22}$$

where G is a deterministic scalar attenuation field; and V is a stationary additive zero mean Gaussian noise RF with variance σ_V^2. The perceived signals in figure 1.6 are defined as follows (see (Sheikh and Bovik, 2006)):

$$E = C + N \quad , \quad F = D + N' \tag{2.23}$$

where N and N' represent stationary white Gaussian noise RFs with variance σ_N^2. If we take the steps outlined in (Sheikh and Bovik, 2006) for VIF index calculation considering scalar GSM, we obtain:

$$I\left(\mathrm{C}^{\mathrm{M}}; \mathrm{E}^{\mathrm{M}} \middle| \mathrm{S}^{\mathrm{M}} = \mathrm{s}^{\mathrm{M}}\right) = I\left(\mathrm{C}^{\mathrm{M}}; \mathrm{E}^{\mathrm{M}} \middle| \mathrm{s}^{\mathrm{M}}\right) = \frac{1}{2}\sum_{i=1}^{\mathrm{M}} log_2\left(\frac{\mathrm{s}_i^2\sigma_U^2 + \sigma_N^2}{\sigma_N^2}\right) \tag{2.24}$$

In the GSM model, the reference image coefficients are assumed to have zero mean. So, for the scalar GSM model, estimates of s_i^2 can be obtained by localized sample variance estimation. The variance σ_U^2 can be assumed to be unity without loss of generality (Sheikh, Bovik and De Veciana, 2005). Thus, Eq. (2.24) is simplified to Eq. (2.25):

$$I\left(\mathrm{C}^{\mathrm{M}}; \mathrm{E}^{\mathrm{M}} \middle| \mathrm{s}^{\mathrm{M}}\right) = \frac{1}{2}\sum_{i=1}^{\mathrm{M}} log_2\left(1 + \frac{\sigma_{\mathrm{c}_i}^2}{\sigma_N^2}\right) \tag{2.25}$$

Similarly, we arrive at Eq. (2.26):

$$I\left(\mathrm{C}^{\mathrm{M}};\mathrm{F}^{\mathrm{M}}\left|\mathrm{s}^{\mathrm{M}}\right.\right)=\frac{1}{2}\sum_{i=1}^{\mathrm{M}}log_2\left(1+\frac{g_i^2\sigma_{\mathrm{c}_i}^2}{\sigma_V^2+\sigma_N^2}\right) \tag{2.26}$$

The final VIF index is defined by Eq. (2.27), as in (Sheikh and Bovik, 2006), but considering a single subband:

$$VIF=\frac{I\left(\mathrm{C}^{\mathrm{M}};\mathrm{F}^{\mathrm{M}}\left|\mathrm{s}^{\mathrm{M}}\right.\right)}{I\left(\mathrm{C}^{\mathrm{M}};\mathrm{E}^{\mathrm{M}}\left|\mathrm{s}^{\mathrm{M}}\right.\right)} \tag{2.27}$$

2.3.2.2 Description of the computational approach

As we explained previously in the SSIM section, we first need to verify the right number of decomposition levels to calculate the proposed $\mathrm{VIF}_{\mathrm{DWT}}$. We perform the experiments on an IVC image database in a similar way to that explained in the SSIM section for the scalar VIF. Figure 2.5 shows the LCC and SRCC between $\mathrm{VIF}_{\mathrm{A}}$ and the MOS values for different decomposition levels. As expected, the $\mathrm{VIF}_{\mathrm{A}}$ prediction accuracy decreases as the number of decomposition levels increases. Therefore, $\mathrm{VIF}_{\mathrm{A}}$ performance is best at N=1. In the second step, we calculate the approximation quality score, $\mathrm{VIF}_{\mathrm{A}}$, between the first-level approximation subbands of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{Y}_{\mathrm{A}}$ (for simplicity, $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{Y}_{\mathrm{A}}$ are used instead of $\mathbf{X}_{\mathrm{A}_1}$ and $\mathbf{Y}_{\mathrm{A}_1}$ here):

$$VIF_A=\frac{\sum_{i=1}^{\mathrm{M}}log_2\left(1+\frac{g_i^2\sigma_{\mathbf{x}_{A,i}}^2}{\sigma_{V_i}^2+\sigma_N^2}\right)}{\sum_{i=1}^{\mathrm{M}}log_2\left(1+\frac{\sigma_{\mathbf{x}_{A,i}}^2}{\sigma_N^2}\right)} \tag{2.28}$$

where M is the number of samples in the approximation subband; $\mathbf{x}_{A,i}$ is the i^{th} image patch in the approximation subband $\mathbf{X}_{\mathrm{A}}$; and $\sigma_{\mathbf{x}_{A,i}}^2$ is the variance of $\mathbf{x}_{A,i}$. The noise variance σ_N^2 is

set to 5 in our approach. The parameters g_i and $\sigma_{V_i}^2$ are estimated as described in (Sheikh and Bovik, 2006), which results in Eq. (2.29) and Eq. (2.30).

$$g_i = \frac{\sigma_{\mathbf{x}_{A,i},\mathbf{y}_{A,i}}}{\sigma_{\mathbf{x}_{A,i}}^2 + \varepsilon} \tag{2.29}$$

where $\sigma_{\mathbf{x}_{A,i},\mathbf{y}_{A,i}}$ is the covariance between image patches $\mathbf{x}_{A,i}$ and $\mathbf{y}_{A,i}$; and ε is a very small constant to avoid instability when $\sigma_{\mathbf{x}_{A,i}}^2$ is zero. In our approach, $\varepsilon = 10^{-20}$.

$$\sigma_{V_i}^2 = \sigma_{\mathbf{y}_{A,i}}^2 - g_i \cdot \sigma_{\mathbf{x}_{A,i},\mathbf{y}_{A,i}} \tag{2.30}$$

All the statistics (the variance and covariance of the image patches) are computed within a local Gaussian square window, which moves (pixel by pixel) over the entire approximation subbands $\mathbf{X}_A$ and $\mathbf{Y}_A$. In this case, a Gaussian sliding window is used in exactly the same way as that defined in step 5 of section 2.2. Because of the smaller resolution of the subbands in the wavelet domain, we can even extract reasonably accurate local statistics with a small, 3×3 sliding window. But, to achieve the best performance and extract accurate local statistics, a larger, 9×9 window is used here. In the simulation section, we show that the VIF_{DWT} can provide accurate scores with the proposed setup.

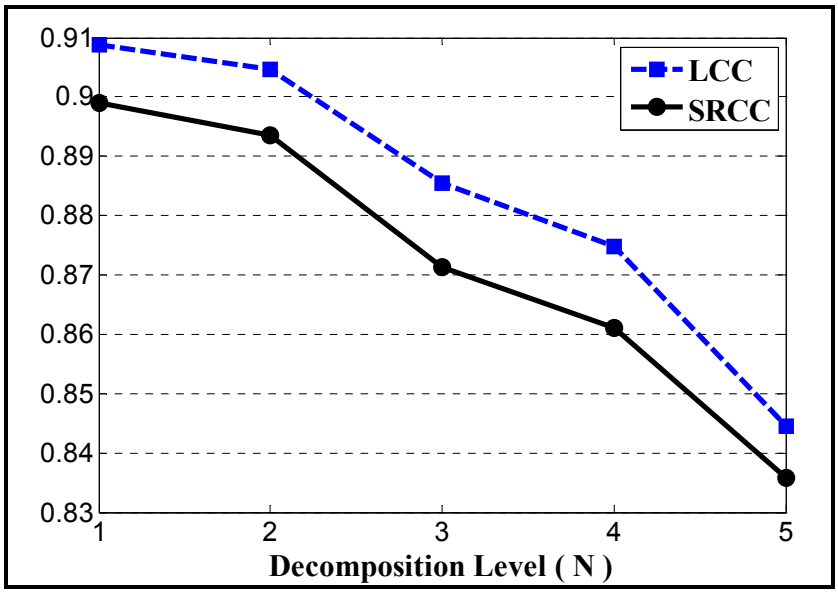

Figure 2.5 LCC and SRCC between the MOS and VIF_A prediction values for various decomposition levels.

In the third step, the edge maps $\mathbf{X}_E$ and $\mathbf{Y}_E$ are computed using Eqs. (2.16) and (2.17). Then, the edge quality score, VIF_E, is calculated between edge maps, as in the second step. Finally, the overall quality measure between images **X** and **Y** is obtained using Eq. (2.14):

$$\begin{aligned} &VIF_{DWT}(\mathbf{X},\mathbf{Y}) = \beta \cdot VIF_A + (1-\beta)\cdot VIF_E \\ &0 < \beta \le 1 \end{aligned} \tag{2.31}$$

where VIF_{DWT} gives the final quality score of images in the range [0,1]. It is worth noting that we skipped the steps for computing a contrast map (Eq. (2.9)) and the pooling procedure as defined in the general framework. That is because the VIF is a non map-based quality score, unlike the SSIM.

2.3.3 PSNR

The conventional PSNR is defined as in Eq. (2.32):

$$\mathrm{PSNR}(\mathbf{X},\mathbf{Y}) = 10 \cdot log_{10}\left(\frac{\mathbf{X}_{max}^2}{\mathrm{MSE}(\mathbf{X},\mathbf{Y})}\right) \tag{2.32}$$

where **X** and **Y** denote the reference and distorted images respectively; $\mathbf{X}_{max}$ is the maximum possible pixel value of the reference image **X** (the minimum pixel value is assumed to be zero). The MSE between **X** and **Y** is calculated as defined in Eq. (1.1). Although the PSNR is still popular because of its ability to easily compute quality in decibels (dB), it cannot adequately reflect the human perception of image fidelity. Other error-based techniques, such as wSNR (Damera-Venkata *et al.*, 2000), NQM (Damera-Venkata *et al.*, 2000), and VSNR (Chandler and Hemami, 2007), are more complex to use, as they follow sophisticated procedures to compute the human visual system (HVS) parameters. In this subsection, we explain how to calculate PSNR-based quality accurately in the discrete wavelet domain using the proposed framework.

The first step is to determine the right number of decomposition levels (N) required to calculate the $\mathrm{PSNR}_{\mathrm{DWT}}$ value. This number can be calculated using Eq. (2.5). To make sure of the validity of Eq. (2.5), we verify the theoretical value by comparing it with the experimental value obtained by performing tests on the IVC database. This database consists of 512×512 images that were subjectively evaluated at a viewing distance of 6 times the screen height.

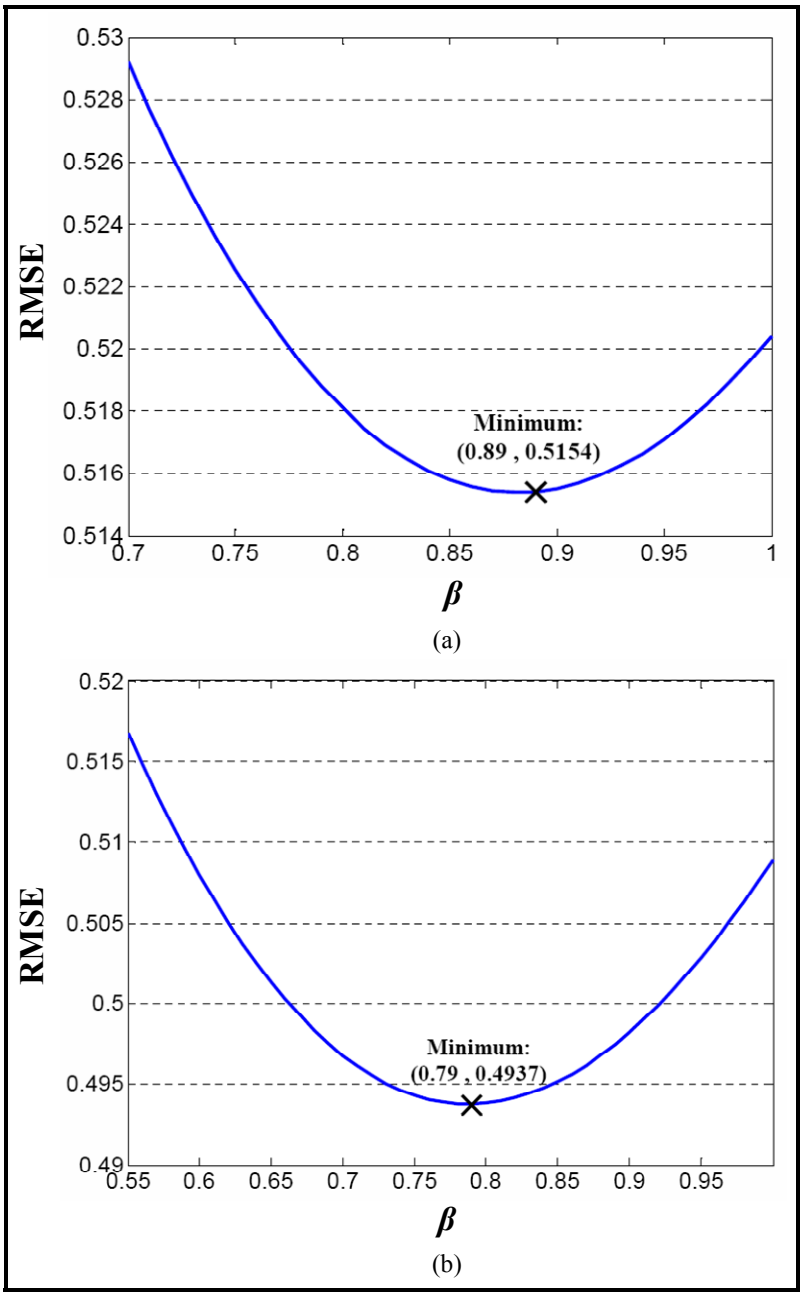

Figure 2.6 RMSE between the MOS and $PSNR_{DWT}$ prediction values for various β values at (a) N=2; (b) N=3.

The plots in figure 2.7 show the LCC and SRCC between $PSNR_A$ and the MOS values for different decomposition levels. It can be seen that $PSNR_A$ attains its best performance at N=3. However, the prediction accuracy for N=2 is very close to that. Based on individual types of distortion, which are available in the corresponding image database, we can

determine which value of N (2 or 3) provides more reliable prediction scores. Table 2.1 lists the SRCC values for four different types of distortion. It is observed that the $PSNR_A$ at N=3 performs better for all types of distortion, expect blurring. When all data (distorted images) are considered, the performance of $PSNR_A$ is superior, at N=3. To reach a fair comparison for $PSNR_{DWT}$, we optimized that full metric with respect to constant β for each decomposition level (N=2 and N=3). When we calculated the root mean square error (RMSE) between the $PSNR_{DWT}$ and MOS values for different β, which reaches its minimum (global) for β = 0.89 at N=2, and for β = 0.79 at N=3 as shown in figure 2.6. Interestingly, these values of β are close to what is suggested in step 7 in the section 2.2, i.e. β is equal to 0.85. The value of β for N=2 is greater than its value for N=3. That is because, for larger N, the resolution of the approximation subband, and consequently the importance of its role in quality prediction, decreases. As table 2.1 shows, the prediction accuracy of $PSNR_{DWT}$ at N=3 is better than that computed at N=2 for all types of distortion. Hence, N=3 is the appropriate decomposition level for the proposed algorithm performing on the IVC database.

The SRCC value for the $PSNR_{DWT}$ at N=3 is 0.9511, which is higher than the value of 0.9368 at N=2. This shows the effectiveness of the proposed edge-map function in improving prediction accuracy, especially for the low-pass filtering type of distortion.

Now, we use Eq. (2.5) to compute the appropriate number of decomposition levels for the IVC database. For that database, k is equal to 6. Thus,

$$\mathrm{N}_{IVC} = \max\left(0, round\left(\log_2\left(512/57.33\right)\right)\right) = 3 \tag{2.33}$$

It can be observed that the theoretical value of N obtained in Eq. (2.33) exactly matches the experimental value explained previously.

We must point out that the LCC has low sensitivity to small variations in β, that is, the proposed β = 0.85 does not drastically affect $PSNR_{DWT}$ performance compared with the optimum β value for the quality prediction across different image databases.

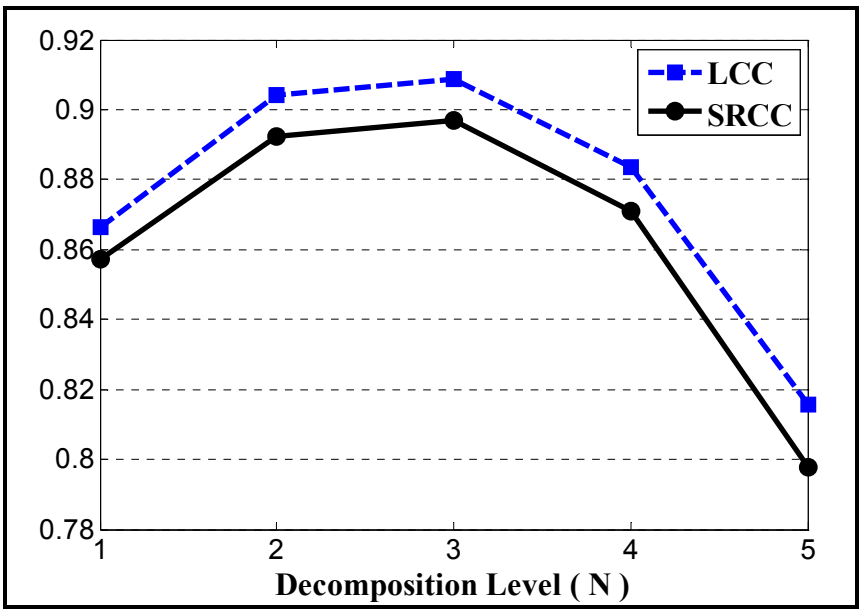

Figure 2.7 LCC and SRCC between the MOS and $PSNR_A$ prediction values for various decomposition levels.

Table 2.1 Values for different types of image distortion in the IVC image database.

Distortion	**$PSNR_A$** (N=2)	**$PSNR_A$** (N=3)	**$PSNR_{DWT}$** (N=2)	**$PSNR_{DWT}$** (N=3)
JPEG	0.8482	0.8699	0.8505	0.865
JPEG2000	0.9210	0.934	0.9262	0.9315
Blur	0.9308	0.9112	0.9368	0.9511
LAR	0.8262	0.8798	0.8668	0.8861
All Data	0.8924	0.8971	0.8964	0.906

In the second step, the edge-map functions of images **X** and **Y** are computed by Eq. (2.6). Then, we calculate the approximation quality score $PSNR_A$ and the edge quality score $PSNR_E$ using Eq. (32), as defined in Eq. (34) and Eq. (35):

$$PSNR_A = PSNR(\mathbf{X}_{A_N}, \mathbf{Y}_{A_N}) \tag{2.34}$$

$$PSNR_E = PSNR(\mathbf{X}_E, \mathbf{Y}_E) \tag{2.35}$$

Finally, the overall quality score $PSNR_{DWT}$ is computed by combining approximation and edge quality scores according to Eq. (2.14):

$$PSNR_{DWT}(\mathbf{X},\mathbf{Y}) = \beta \cdot PSNR_{A} + (1-\beta) \cdot PSNR_{E} \quad , \quad 0 < \beta \leq 1 \tag{2.36}$$

where $PSNR_{DWT}$ gives the final quality score of the images in dB.

2.3.4 Absolute difference (AD)

To verify the performance of our framework more generally, we investigate how it works if the AD of the images is considered as the IQM. As in previous cases, we first need to know the required number of decomposition levels in order to calculate the AD_{DWT} value. When we perform a test on the IVC image database in the same way as before, figure 2.8 is obtained, which shows the LCC and SRCC between the mean AD_A and MOS values for different decomposition levels. Like the $PSNR_A$, the AD_A performs well at N=2 and N=3. Table 2.2 shows SRCC values between the AD-based metrics and MOS values for two and three decomposition levels. According to table 2.2, the performances of the AD_A and AD_{DWT} at N=3 are better than N=2 for individual types of distortion. Like the computation of $PSNR_{DWT}$ in the previous subsection, we optimized AD_{DWT} with respect to constant β, which results in $\beta = 0.89$ for N = 2 and $\beta = 0.72$ for N = 3. When the value of N is calculated by Eq. (2.5), the result is three levels of decomposition for the IVC database that match the experimental value.

In the second step, we calculate the approximation AD map, AD_A, between the approximation subbands of **X** and **Y**.

$$AD_{A}(m,n) = \left| \mathbf{X}_{A_N}(m,n) - \mathbf{Y}_{A_N}(m,n) \right| \tag{2.37}$$

Where (m,n) shows a sample position in the approximation subband.

In the third step, the edge-map function images **X** and **Y** are defined in Eqs. (2.6),(2.7), and the edge AD map, AD_E, is calculated between the edge maps $\mathbf{X}_E$ and $\mathbf{Y}_E$ in the next step.

$$\mathrm{AD}_{\mathrm{E}}(m,n) = \left|\mathbf{X}_{\mathrm{E}}(m,n) - \mathbf{Y}_{\mathrm{E}}(m,n)\right| \tag{2.38}$$

In the fourth step, the contrast map is obtained using Eq. (2.9), and then AD_A and AD_E are pooled using the contrast map to calculate the approximation and edge quality scores S_A and S_E.

$$S_A = \frac{\sum_{j=1}^{\mathrm{M}} Contrast(m,n)\cdot \mathrm{AD}_{\mathrm{A}}(m,n)}{\sum_{j=1}^{\mathrm{M}} Contrast(m,n)} \tag{2.39}$$

$$S_E = \frac{\sum_{j=1}^{\mathrm{M}} Contrast(m,n)\cdot \mathrm{AD}_{\mathrm{E}}(m,n)}{\sum_{j=1}^{\mathrm{M}} Contrast(m,n)} \tag{2.40}$$

The final quality score, AD_{DWT}, is calculated using Eq. (2.41).

$$\mathrm{AD}_{\mathrm{DWT}}(\mathbf{X},\mathbf{Y}) = \beta\cdot S_A + (1-\beta)\cdot S_E \tag{2.41}$$

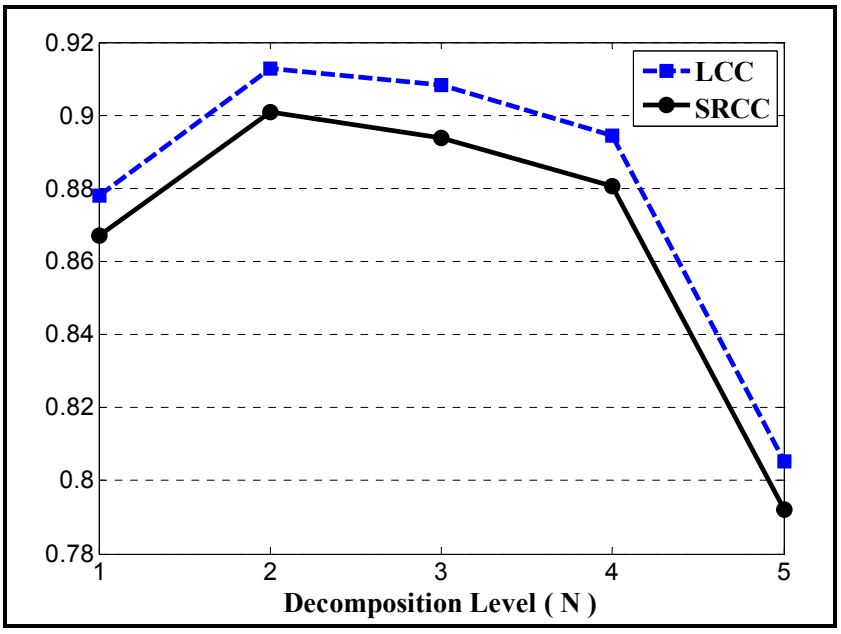

Figure 2.8 LCC and SRCC between the MOS and mean AD_A prediction values for various decomposition levels.

Table 2.2 SRCC values for different types of image distortion in the IVC image database.

Distortion	**AD_A** (N=2)	**AD_A** (N=3)	**AD_{DWT}** (N=2)	**AD_{DWT}** (N=3)
JPEG	0.8965	0.9101	0.8851	0.9013
JPEG2000	0.9294	0.9348	0.9286	0.9283
Blur	0.9157	0.9142	0.9323	0.9368
LAR	0.8543	0.8885	0.8897	0.9015
All Data	0.9076	0.9072	0.9125	0.9233

CHAPTER 3

THE PERFORMANCE EVALUATION OF THE PROPOSED VISUAL QUALITY ASSESSMENT FRAMEWORK

In this chapter, we investigate the performance of our proposed framework from different aspects. First, the computational complexity of our approach is discussed. Then, the prediction accuracy of our methods is verified, and compared with other well-known methods for images. Finally, we bring the results to show performance of our approach tested on video sequences.

3.1 Computational complexity of the algorithms

In spite of the number of steps required to calculate the final quality score, the computational complexity of the proposed algorithms is low. Here, we discuss various different aspects of the complexity of the approach. The resolution of the approximation subband and edge map is a quarter of that of the original image. Lower resolutions mean that fewer computations are required to obtain the image statistics or quality maps, e.g. SSIM maps for the $SSIM_{DWT}$. Because of the smaller resolution of the subbands in the wavelet domain, we can extract accurate local statistics with a smaller sliding window. For example, the spatial SSIM in (Wang *et al.*, 2004) uses an 11×11 window by default, while in the next section we show that the $SSIM_{DWT}$ can provide accurate scores with a 4×4 window. A smaller window reduces the number of computations required to obtain local statistics. As can be seen from Eq. (2.9), the local statistics calculated for Eq. (2.15) and Eq. (2.18) are used to form the contrast map. Therefore, in computing $SSIM_{DWT}$, the contrast map does not impose a large computational burden.

Probably the most complex part of the approach is wavelet decomposition. A simple wavelet can be used to reduce complexity. We used the Haar wavelet for image decomposition. As this wavelet has the shortest filter length, it makes the filtering process simpler. The use of the Haar wavelet makes it possible to calculate VIF_{DWT} using the scalar GSM model with a

complexity of about 5% of the original VIF index, which uses an over-complete steerable pyramid transform. Furthermore, the use of the Haar wavelet enables us to investigate the computational complexity of simple algorithms, e.g. $PSNR_{DWT}$, mathematically and compare them to other methods like PSNR, as explained below.

To calculate the PSNR between two images, we need 1 subtraction, 1 multiplication (square), and 1 addition for every input pixel. Therefore, this calculation requires 3 operations per input pixel.

In the decomposition stage, one operation per input pixel must be performed to obtain a desired image subband using the Haar wavelet. That is because the Haar wavelet is actually a simple averaging. For example, to obtain the second level approximation subband, we need to perform 15 additions and 1 shift (as a division) for every 4×4=16 neighboring pixels, which results in 1 operation per input pixel. As we apply the DWT to both the reference and the distorted images, we need $(2+3/(4^N))$ operations per input pixel to calculate $PSNR_A$ (with N-level wavelet decomposition). Since N is greater than or equal to unity ($N \geq 1$), the computational complexity of $PSNR_A$ is less than that of the PSNR. However, in the next section, we show that $PSNR_A$ is much more accurate than the PSNR in predicting quality scores.

In order to calculate $PSNR_E$, we first need to compute edge maps for the reference and distorted images. By analyzing Eq. (2.6), Eq. (2.7), and Eq. (2.35), we find that the number of operations per input pixel for calculating $PSNR_E$ is found as in Eq. (3.1) (considering the square root as s operations):

$$\#\,of\ \ operations\ per\ input\ pixel\ \ (\mathrm{PSNR_E}) = 2N\cdot\left(3+(8+s)/4^N\right)+\frac{3}{4^N}=3\cdot\left(2N+\left(1+\frac{1}{3}N(16+2s)\right)/4^N\right) \tag{3.1}$$

The value of s is about 30 for Intel processor architectures (*Intel® 64 and IA32 architectures optimization reference manual*, April 2012). For instance, for an N of 2, the complexity of $PSNR_E$ is about 7.2 times that of the PSNR (and about 7 times greater for N=3). A comparison based on a C++ implementation of SSIM shows that, for a 640×480 image, it is approximately 115 times more complex than the PSNR. Thus, calculation of $PSNR_{DWT}$ is not computationally expensive relative to other metrics. However, we should mention that $PSNR_A$ alone gives very accurate scores, while $PSNR_E$ is most effective when taking into account certain distortions, like fast fading channel distortion.

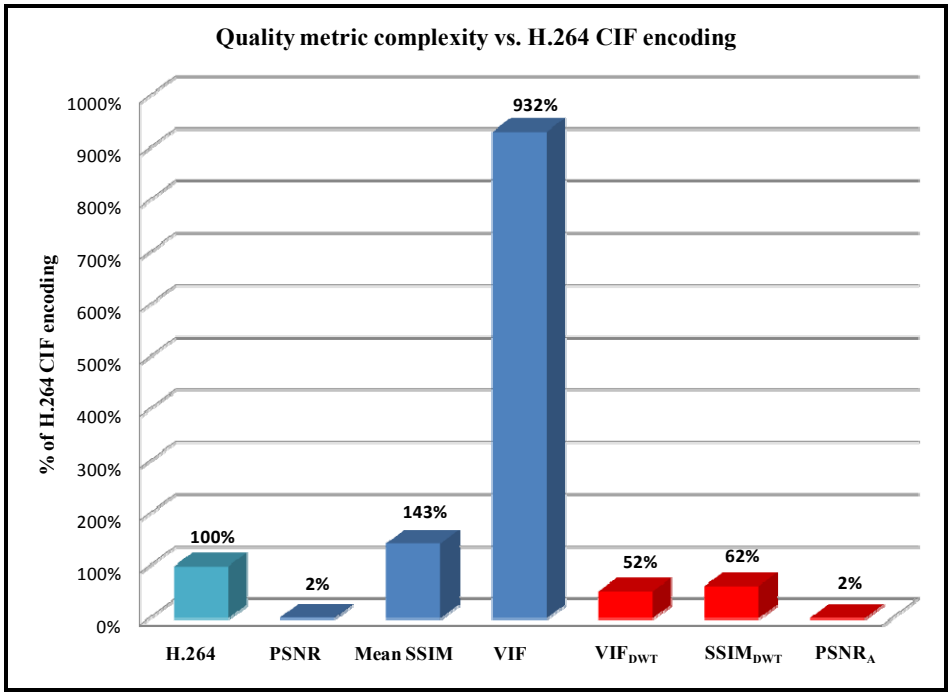

Figure 3.1 Comparison of the complexity of various quality metrics vs. H.264 encoding complexity.

In order to develop a practical concept of the complexity of the various metrics, we chose the complexity of IPP-based H.264 baseline encoding (Intel® Integrated Performance Primitives (Intel® IPP)) for CIF-sized videos as the benchmark and compared it to the complexity of some of the metrics. We used Intel IPP encoder instead of JM reference software because it is a highly optimized for microprocessors and is more representative of real-life products. In order to verify the computational complexity of those metrics, we measured the execution

time of the algorithm based on the elapsed CPU time. Figure 3.1 shows the bar graph of quality metric complexity versus H.264 CIF encoding. We used C/C++ implementations for timing measurement. As can be observed from figure 3.1, the computational complexity of PSNR and $PSNR_A$ is about 2% of the H.264 encoding. In the next section, we show that $PSNR_A$ prediction accuracy is much greater than that of conventional PSNR.

3.2 Verification of quality prediction accuracy of metrics for images

In the previous chapter, we used the IVC image database for some verification with respect to decomposition levels. In this section, the performance of the proposed algorithm for the quality calculation is evaluated on two different image database: the LIVE Image Quality Assessment Database, Release 2 (Sheikh *et al.*) and the Tampere Image Database 2008 (TID2008), version 1.0 (Ponomarenko *et al.*, 2008).

At the first stage, we bring the results obtained from the LIVE database. This database consists of 779 distorted images derived from 29 original color images using five types of distortion: JPEG compression, JPEG2000 compression, Gaussian white noise (GWN), Gaussian blurring (GBlur), and the Rayleigh fast fading (FF) channel model. The realigned subjective quality data for the database are used in all experiments (Sheikh *et al.*).

In this section, four performance metrics are applied, in addition to the statistical *F* test, to measure the performance of the objective models. The first metric is the Pearson correlation coefficient (LCC) between the Difference Mean Opinion Score (DMOS) and the objective model outputs after nonlinear regression. The correlation coefficient gives an evaluation of prediction accuracy. We use the five-parameter logistical function defined in (Sheikh, Sabir and Bovik, 2006) for nonlinear regression. The second metric is the Spearman rank correlation coefficient (SRCC), which provides a measure of prediction monotonicity. The third metric is root mean square error (RMSE), which is considered as a measure of prediction consistency. The fourth metric is the Kendall rank correlation coefficient (KRCC), which is used to measure the association or statistical dependence between two measured

quantities in cases where the population distribution of either or both variables is unknown, that is, it is a measure of the degree of correspondence between sets of rankings.

In order to put the performance evaluation of our proposed scheme into the proper context, we compare our quality assessment algorithms with other quality metrics, including the conventional PSNR, the spatial domain mean SSIM (Wang *et al.*, 2004), an autoscale version of SSIM that performs downsampling on images (Wang), and a weighted SNR (wSNR) (Damera-Venkata *et al.*, 2000), in which the images are filtered by the CSF specified in (Mannos and Sakrison, 1974) and (Mitsa and Varkur, 1993). For the LIVE database, we set k at Eq. (2.5) equal to 3, based on the experimental setup and the decomposition level calculated for each image using Eq. (2.5).

Tables 3.1, 3.2, 3.3 and 3.4 show the results of performance metrics (LCC, SRCC, RMSE and KRCC) for each type of image distortion in the LIVE database. To understand the effect of the contrast map (Eq. (2.9)) in improving quality prediction, we can compare the results of the mean($SSIM_A$) and mean(AD_A) rows with $S_A(SSIM_A)$ and $S_A(AD_A)$ rows in the corresponding tables. $S_A(SSIM_A)$ and $S_A(AD_A)$ are showing quality scores after weighing the quality map $SSIM_A$ and AD_A by the proposed contrast map. It is observed that the SRCC of the mean($SSIM_A$) increases from 0.9441 to 0.9573 for $S_A(SSIM_A)$, which corresponds to a significant improvement. The SRCC of mean(AD_A) increases from 0.9351 to 0.9421 for $S_A(AD_A)$, which shows improvement due to application of the contrast map.

The performance of $SSIM_{DWT}$ is the best of all the structural metrics. For SNR-based metrics, the SRCC of $PSNR_A$ is 0.9307, which is higher than that of conventional PSNR (0.8756) and even wSNR (0.9240), while its complexity is lower than that of conventional PSNR. The performance of $PSNR_{DWT}$ is better than $PSNR_A$ for GWN, GBlur, and FF types of distortion.

Table 3.1 LCC values after nonlinear regression for the LIVE image database.

Model	JPEG	JPEG2000	GWN	GBlur	FF	All Data
$\mathbf{SSIM_{spatial}}$	0.9504	0.9413	0.9747	0.8743	0.9449	0.9038
$\mathbf{SSIM_{autoscale}}$	0.9778	0.9669	0.9808	0.9483	0.9545	0.9446
$\mathbf{mean(SSIM_A)}$	0.9762	0.9699	0.9645	0.9548	0.9625	0.9412
$\boldsymbol{S_A(SSIM_A)}$	0.9782	0.9705	0.9724	0.9724	0.9730	0.9534
$\mathbf{SSIM_{DWT}}$	0.9835	0.9747	0.9791	0.9690	0.9735	0.9556
$\mathbf{PSNR_{spatial}}$	0.8879	0.8996	0.9852	0.7835	0.8895	0.8701
wSNR	0.9692	0.9351	0.9776	0.9343	0.8983	0.9211
$\boldsymbol{S_A(PSNR_A)}$	0.9793	0.9542	0.9806	0.9241	0.8868	0.9288
$\mathbf{PSNR_{DWT}}$	0.9787	0.9549	0.9838	0.9234	0.8994	0.9300
mean(ADA)	0.9787	0.9425	0.9587	0.9083	0.8748	0.9296
$\boldsymbol{S_A(AD_A)}$	0.9817	0.9587	0.9637	0.9307	0.9005	0.9350
$\mathbf{AD_{DWT}}$	0.9807	0.9579	0.9678	0.9258	0.9064	0.9344
VIF	0.9864	0.9773	0.9901	0.9742	0.9677	0.9593
$\boldsymbol{S_A(VIF_A)}$	0.9856	0.9735	0.9904	0.9615	0.9611	0.9639
$\mathbf{VIF_{DWT}}$	0.9852	0.9740	0.9906	0.9652	0.9650	0.9654

Table 3.2 SRCC values after nonlinear regression for the LIVE image database.

Model	JPEG	JPEG2000	GWN	GBlur	FF	All Data
$\mathbf{SSIM_{spatial}}$	0.9449	0.9355	0.9629	0.8944	0.9413	0.9104
$\mathbf{SSIM_{autoscale}}$	0.9764	0.9614	0.9694	0.9517	0.9556	0.9479
$\mathbf{mean(SSIM_A)}$	0.9738	0.9634	0.9490	0.9620	0.9622	0.9441
$\boldsymbol{S_A(SSIM_A)}$	0.9779	0.9634	0.9577	0.9703	0.9699	0.9573
$\mathbf{SSIM_{DWT}}$	0.9819	0.9678	0.9683	0.9707	0.9708	0.9603
$\mathbf{PSNR_{spatial}}$	0.8809	0.8954	0.9854	0.7823	0.8907	0.8756
wSNR	0.9610	0.9292	0.9749	0.9330	0.8990	0.9240
$\boldsymbol{S_A(PSNR_A)}$	0.9647	0.9499	0.9777	0.9219	0.8853	0.9307
$\mathbf{PSNR_{DWT}}$	0.9648	0.9494	0.9818	0.9230	0.9004	0.9325
mean(ADA)	0.9654	0.9393	0.9757	0.9056	0.8819	0.9351
$\boldsymbol{S_A(AD_A)}$	0.9666	0.9553	0.9805	0.9335	0.9067	0.9421
$\mathbf{AD_{DWT}}$	0.9661	0.9546	0.9835	0.9290	0.9131	0.9412
VIF	0.9845	0.9696	0.9858	0.9726	0.9649	0.9635
$\boldsymbol{S_A(VIF_A)}$	0.9837	0.9669	0.9848	0.9618	0.9597	0.9663
$\mathbf{VIF_{DWT}}$	0.9829	0.9680	0.9853	0.9657	0.9641	0.9681

Table 3.3 RMSE values after nonlinear regression for the LIVE image database.

Model	JPEG	JPEG2000	GWN	GBlur	FF	All Data
$\mathbf{SSIM_{spatial}}$	9.9106	8.5151	6.2603	8.9663	9.3253	11.6907
$\mathbf{SSIM_{autoscale}}$	6.6753	6.4368	5.4740	5.8625	8.4956	8.9673
$\mathbf{mean(SSIM_A)}$	6.9082	6.1465	7.3837	5.4910	7.7287	9.2270
$\boldsymbol{S_A(SSIM_A)}$	6.6122	6.0858	6.5263	4.3060	6.5745	8.2438
$\mathbf{SSIM_{DWT}}$	5.7631	5.6426	5.7348	4.5645	6.5199	8.0480
$\mathbf{PSNR_{spatial}}$	14.6532	11.0174	4.7918	11.4784	13.0148	13.4685
wSNR	7.8400	8.9401	5.8895	6.5843	12.5139	10.6353
$\boldsymbol{S_A(PSNR_A)}$	6.4439	7.5467	5.4853	7.0571	13.1668	10.1224
$\mathbf{PSNR_{DWT}}$	6.5396	7.4949	5.0148	7.0898	12.4549	10.0409
mean(ADA)	6.5429	8.4337	7.9597	7.7268	13.8038	10.0702
$\boldsymbol{S_A(AD_A)}$	6.0741	7.1794	7.4926	6.7566	12.3942	9.6890
$\mathbf{AD_{DWT}}$	6.2308	7.2456	7.0510	6.9832	12.0362	9.7349
VIF	5.2420	5.3498	3.9186	4.1669	7.1968	7.7122
$\boldsymbol{S_A(VIF_A)}$	5.3883	5.7682	3.8822	5.0783	7.8757	7.2755
$\mathbf{VIF_{DWT}}$	5.4685	5.7125	3.8261	4.8313	7.4785	7.1284

Table 3.4 KRCC values after nonlinear regression for the LIVE image database.

Model	JPEG	JPEG2000	GWN	GBlur	FF	All Data
$\mathbf{SSIM_{spatial}}$	0.7933	0.7694	0.8364	0.7136	0.7824	0.7311
$\mathbf{SSIM_{autoscale}}$	0.8650	0.8239	0.8523	0.8010	0.8207	0.7963
$\mathbf{mean(SSIM_A)}$	0.8589	0.8273	0.8013	0.8247	0.8320	0.7873
$\boldsymbol{S_A(SSIM_A)}$	0.8701	0.8267	0.8247	0.8498	0.8489	0.8147
$\mathbf{SSIM_{DWT}}$	0.8850	0.8381	0.8511	0.8483	0.8513	0.8219
$\mathbf{PSNR_{spatial}}$	0.6912	0.7106	0.8939	0.5847	0.7069	0.6865
wSNR	0.8253	0.7613	0.8554	0.7780	0.7293	0.7613
$\boldsymbol{S_A(PSNR_A)}$	0.8334	0.7994	0.8669	0.7513	0.7050	0.7723
$\mathbf{PSNR_{DWT}}$	0.8344	0.7991	0.8808	0.7534	0.7228	0.7731
mean(ADA)	0.8370	0.7798	0.8615	0.7280	0.7077	0.7785
$\boldsymbol{S_A(AD_A)}$	0.8415	0.8128	0.8759	0.7726	0.7389	0.7916
$\mathbf{AD_{DWT}}$	0.8380	0.8115	0.8868	0.7626	0.7479	0.7883
VIF	0.8943	0.8474	0.8981	0.8584	0.8389	0.8278
$\boldsymbol{S_A(VIF_A)}$	0.8928	0.8360	0.8960	0.8299	0.8312	0.8365
$\mathbf{VIF_{DWT}}$	0.8898	0.8407	0.8969	0.8372	0.8389	0.8402

To verify the validity of the proposed framework, we check the performance of AD_{DWT}. Table 3.2 shows that its performance is close to mean $SSIM_A$ and $SSIM_{autoscale}$. The SRCC value for VIF_{DWT} is 0.9681, which is higher than the SRCC value of the VIF index (0.9635) defined in (Sheikh and Bovik, 2006). VIF_{DWT} has the best performance of all the IQMs we describe here.

To assess the statistical significance of each metric's performance relative to that of other metrics, a two-tailed *F* test was performed on the residual differences between the IQM predictions and the DMOS. The *F* test is used to determine whether one metric has significantly larger residuals (greater prediction error) than another (Video quality experts group, 2003). The *F* statistic is defined by a ratio of variances of prediction errors (residuals) from two IQMs. The more this ratio deviates from 1, the stronger the evidence for unequal population variances. Values of $F > F_{critical}$ or $F < 1/F_{critical}$ indicate that residuals resulting from one quality metric are statistically distinguishable from the residuals of another quality metric (i.e., significantly larger or smaller). $F_{critical}$ is computed based on the number of residuals and a significance level of α. In this paper, we used $\alpha = 0.05$, which results in $F_{critical} = 1.151$. Table 3.5 shows the *F* statistic obtained based on the residuals from each structurally based metric against the residuals from $SSIM_{autoscale}$. It is observed that $SSIM_{DWT}$ has significantly smaller residuals than $SSIM_{autoscale}$, and $SSIM_{spatial}$ has significantly larger residuals than $SSIM_{autoscale}$. Results in table 3.6 show that the proposed VIF_{DWT} outperforms the VIF index when all the distorted images are considered. *F* statistic values for error-based metrics are presented in table 3.7. As can be seen, the performance of AD_{DWT} is statistically superior to that of the other error-based models. Bold values in tables 3.5, 3.6, and 3.7 are statistically different from other values.

Figure 3.2 shows the scatter plots of DMOS vs. various quality metrics for all the distorted images. It is evident that the $SSIM_{DWT}$, $PSNR_{DWT}$, and VIF_{DWT} predictions are more linear and more uniformly scattered compared to other models.

Table 3.5 *F*-test results on the residual error predictions of different structure-based IQMS.

Model	Residual Variance	*F*-statistic
$SSIM_{spatial}$	136.8492	**1.7002**
$SSIM_{autoscale}$	80.4888	1.0000
mean $SSIM_A$	85.2478	1.0591
S_A	68.0476	**0.8454**
$SSIM_{DWT}$	64.8528	**0.8057**

Table 3.6 *F*-test results on the residual error predictions of various information-theoretic-based IQMS.

Model	Residual Variance	*F*-statistic
VIF	59.5548	1.0000
VIF_A	52.9965	0.8899
VIF_{DWT}	50.8794	**0.8543**

Table 3.7 *F*-test results on the residual error predictions of various error-based IQMS.

Model	Residual Variance	*F*-statistic
$PSNR_{spatial}$	181.6336	1.6038
wSNR	113.2543	1.0000
$PSNR_A$	102.5939	0.9059
$PSNT_{DWT}$	100.9495	0.8914
AD_{DWT}	94.8892	**0.8378**

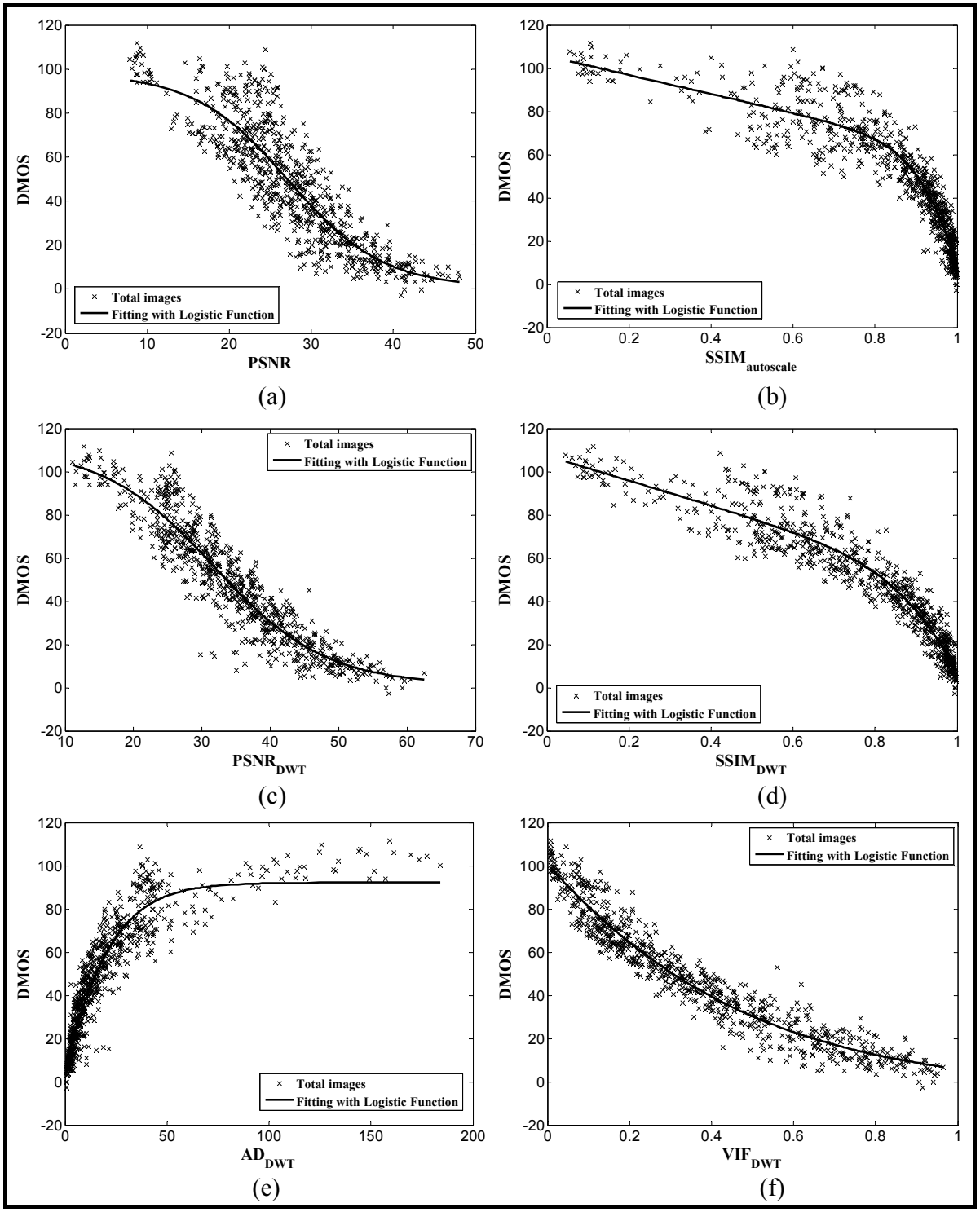

Figure 3.2 Scatter plots of DMOS versus model prediction for all distorted images in the LIVE database. (a) PSNR; (b) $SSIM_{autoscale}$; (c) $PSNR_{DWT}$; (d) $SSIM_{DWT}$; (e) AD_{DWT}; (f) VIF_{DWT}.

Table 3.8 Performance comparison of image quality assessment models for TID2008 image database (only images with distortion types of additive Gaussian noise, Gaussian blur, JPEG compression, JPEG2000 compression, JPEG transmission errors, and JPEG2000 transmission errors are included).

Model	LCC	SRCC	RMSE	KRCC
$\mathbf{SSIM_{spatial}}$	0.7506	0.7775	0.9581	0.5735
$\mathbf{SSIM_{Autoscale}}$	0.8571	0.8822	0.7469	0.6855
$\mathbf{S_A(SSIM_A)}$	0.8789	0.8870	0.6917	0.6997
$\mathbf{SSIM_{DWT}}$	0.8845	0.8933	0.6765	0.7092
$\mathbf{PSNR_{spatial}}$	0.7822	0.8038	0.9034	0.6053
wSNR	0.8603	0.8791	0.7393	0.6960
$\boldsymbol{S_A(PSNR_A)}$	0.8911	0.9097	0.6581	0.7446
$\mathbf{PSNR_{DWT}}$	0.8931	0.9127	0.6522	0.7454
$\mathbf{S_A(AD_A)}$	0.9210	0.9274	0.5648	0.7712
$\mathbf{AD_{DWT}}$	0.9202	**0.9304**	0.5676	0.7758
VIF	0.9081	0.8951	0.6072	0.7171
$\boldsymbol{S_A(VIF_A)}$	0.9001	0.8981	0.6317	0.7222
$\mathbf{VIF_{DWT}}$	0.9002	0.8929	0.6314	0.7155

In the second stage of our simulations, we present the results obtained for the TID2008 image database. The TID2008 contains 25 reference images and 1,700 distorted images, including 17 types of distortions and 4 levels of distortions for each of the reference images (Ponomarenko *et al.*, 2008). Some distortions of this database are applied on color components of images. Since all the aforementioned metrics work for grayscale images and do not take into consideration color information, we only consider the following types of distortions: additive Gaussian noise, Gaussian blur, JPEG compression, JPEG2000 compression, JPEG transmission errors, and JPEG2000 transmission errors. These distortions are nearly the same as distortions on the LIVE image database.

Table 3.8 shows the results of performance metrics tested on the TID2008 image database for the mentioned distortion types. According to table 3.8, it is observed that the SRCC of AD_{DWT} is 0.9304 which is the highest among all metrics. The performance of $PSNR_{DWT}$ is better than both $SSIM_{DWT}$ and VIF. Results on table 3.8 confirm the validity of our

framework. It also shows that the performance of metrics can change noticeably between different image databases.

3.3 Verification of quality prediction accuracy of metrics for videos

Although our visual quality assessment framework does not provide any temporal pooling strategy, like (Wang, Lu and Bovik, 2004), for quality prediction of video sequences, we expect our metrics to perform well for video compression, where distortions are uniformly dispersed spatially and temporally.

Therefore, to more effectively verify the accuracy of the proposed algorithms for quality calculation and also the correctness of the formula defined for deciding on N (Eq. (2.5)), we tested the performance of our algorithm on the LIVE video quality database (Seshadrinathan *et al.*, 2010a),(Seshadrinathan *et al.*, 2010b). To simulate our algorithm, the parameter N was computed using Eq. (2.5) and β was used as before, i.e. $\beta = 0.85$. Table 3.9 lists the results of the performance evaluation on the LIVE video database.

Table 3.9 Performance comparison of image quality assessment models for H.264/AVC video compression using the LIVE video quality database.

Model	LCC	SRCC	RMSE	KRCC
$\mathbf{SSIM_{spatial}}$	0.6878	0.6561	7.884	0.4897
$\mathbf{SSIM_{Autoscale}}$	0.7445	0.7129	7.248	0.5564
$\mathbf{S_A(SSIM_A)}$	0.7163	0.7021	7.575	0.5410
$\mathbf{SSIM_{DWT}}$	0.7367	**0.7278**	7.341	**0.5744**
$\mathbf{PSNR_{spatial}}$	0.5845	0.4730	8.808	0.3436
wSNR	0.5990	0.5248	8.692	0.3641
$\boldsymbol{S_A(PSNR_A)}$	0.6265	0.6008	8.460	0.4333
$\mathbf{PSNR_{DWT}}$	0.6435	**0.6420**	8.309	**0.4769**
$\mathbf{S_A(AD_A)}$	0.6322	0.6036	8.410	0.4333
$\mathbf{AD_{DWT}}$	0.6358	0.6066	8.379	0.4385
VIF	0.6501	0.6313	8.249	0.4692
$\boldsymbol{S_A(VIF_A)}$	0.7024	0.6814	7.727	0.5077
$\mathbf{VIF_{DWT}}$	0.7022	**0.6732**	7.729	**0.5026**

The results in table 3.9 confirm that the proposed framework is efficient in providing accurate visual quality scores. The $SSIM_{DWT}$ achieves the best performance, with the SRCC of 0.7278, among all metrics. The $PSNR_{DWT}$ performs much better than the conventional PSNR, and even provides better quality prediction than the VIF. It is also observed that the proposed VIF_{DWT} outperforms the original VIF index.

3.4 Conclusion

In this thesis, we proposed a novel framework for calculating quality prediction scores in the discrete wavelet domain using the Haar wavelet. The proposed methods and formulas were described and verified on the IVC image database as a training database, and finally validated on the LIVE image database, the TID2008 image database, and the LIVE video database as test image and video databases. Our results show that the approximation subband of decomposed images plays an important role in improving quality assessment performance, and also in complexity reduction. To compute the map-based metrics, we defined a contrast map, which takes advantage of basic HVS characteristics for discrete wavelet domain pooling of quality maps. Also, we defined an edge map for each image to represent an estimate of image edges. Although computing the edge map increases the metrics' complexity compared to just considering approximation bands, the complexity is still very low compared to other metrics and increases the accuracy. Moreover, based on the application we can decide to use or not the edges. We described four different quality assessment methods using the framework, including $SSIM_{DWT}$, VIF_{DWT}, $PSNR_{DWT}$, and AD_{DWT}.

Simulation results on the LIVE image database show that the proposed VIF_{DWT} is slightly more accurate than the original VIF index, while its complexity is much lower. Also, we described $PSNR_{DWT}$, which gives the quality in decibels (dB). $PSNR_{DWT}$ is more accurate than conventional PSNR, while its complexity is very low compared to that of other metrics. We also proposed AD_{DWT} to verify the validity of our general framework. For error-based metrics, a formula was proposed to compute the appropriate level of wavelet decomposition

at the desired viewing distance. It is notable that there are not many parameters for calculating metrics in our framework. Calculating the approximation quality score needs no extra parameters in addition to the original versions of the metrics. Also the same parameters used across the databases and no database-dependant optimizations were performed on parameters. Since the proposed framework provides a way to calculate quality with very good tradeoffs between accuracy and complexity, it can be used efficiently in wavelet-based image/video processing applications.

CHAPTER 4

MODE DECISION IN H.264 VIDEO ENCODING CONSIDERING THE HUMAN VISUAL SYSTEM

4.1 Background and related works

One of the latest standards in video coding is called H.264 (also called MPEG-4 AVC, Advanced Video Coding defined in MPEG-4 Part 10). The H.264/AVC video coding standard has significantly improved coding efficiency, compared to older video codecs such as MPEG-2 and H.263, at the expense of higher computational complexity. Software-based studies on this standard show that H.264 offers up to 50% better compression than MPEG-2 and up to 30% better than H.263+ and MPEG-4 advanced simple profile (Wiegand *et al.*, 2003b). H.264 has been adopted by many application standards such as Blu-ray, HD DVD, DVB-H, HD-DTV.

Since, the optimal coding decisions are usually very computationally expensive to be practical, the video encoder has to manage a trade-off between quality and complexity. For this reason, sub-optimal techniques are normally used for video encoding. However, these techniques must be satisfactory in terms of the encoding speed and decoded video quality.
After motion estimation, the second most computationally expensive phase in the encoding process is the macroblock mode decision during inter-frame and intra-frame predictions. These decisions have direct impacts on the resulting video quality.

4.1.1 Inter prediction and macroblock partitions in H.264

In H.264 inter prediction a block of luminance (luma) and chrominance (chroma) samples is predicted from a frame that has previously been coded, a reference picture. The block of samples or pixels to be predicted (a macroblock or sub-macroblock partition) can vary in size from a complete macroblock, i.e. 16 × 16 luma samples and corresponding chroma samples, down to a 4×4 block of luma samples and corresponding chroma samples (Richardson,

2010). Inter prediction uses previously coded pictures, stored in a decoded picture buffer (DPB), as reference pictures. For example in P slices, all reference pictures in the DPB are indexed as a single list (List0) and the first picture in the list is by default the most recently decoded picture.

Each 16×16 P or B macroblock is predicted using a range of block sizes. The macroblock may be split up into 4 kinds of macroblock partitions: one 16×16 partition, two 8×16 partitions, two 16×8 partitions, and four 8×8 partitions. If 8×8 partitions are chosen, then each 8×8 sub-macroblock of luma samples and its associated chroma samples, can be independently partitioned into one 8×8 block, two 4×8 blocks, two 8×4 blocks or four 4×4 blocks. Figure 4.1(a) demonstrates the splitting process for a macroblock partitions, and figure 4.1(b) shows the process for sub-macroblock partitions.

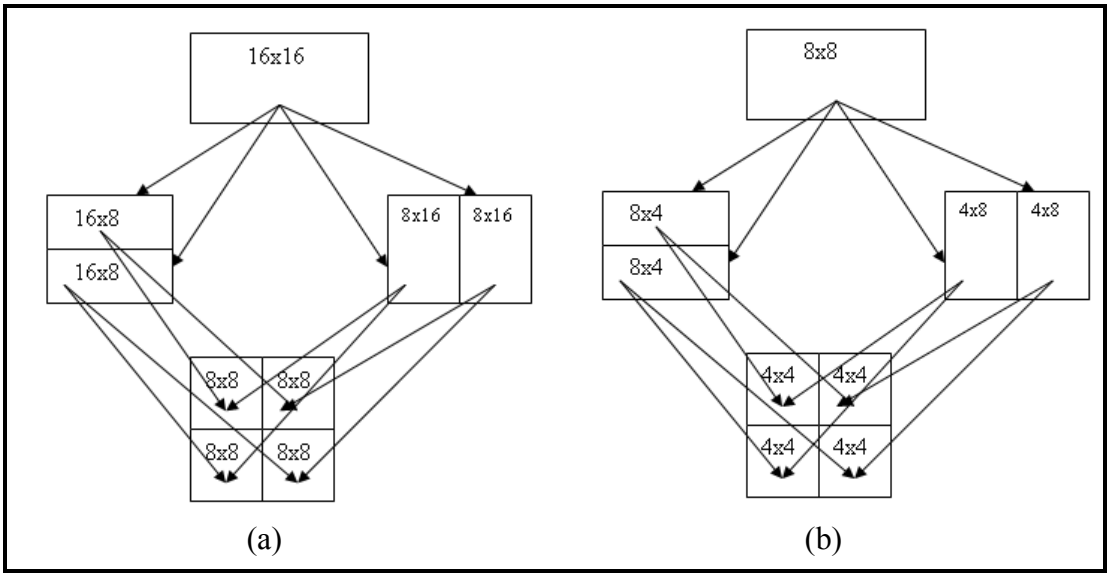

Figure 4.1 Macroblock splitting process in H.264; (a) macroblock partitions; (b) sub-macroblock partitions.

Each macroblock partition and sub-macroblock partition has one or two motion vectors pointing to an area of the same size in the reference frame. In a P macroblock, we have one motion vector per macroblock partition or sub-macroblock partition. However, a partition in a B macroblock can have one or two reference frames and consequently one or two motion

vectors. It is noteworthy that prediction can be performed independently for each macroblock partition, but the coding is still based on a 4×4 block.

4.1.2 Rate distortion optimized mode selection in H.264

In this subsection, we first explain the macroblock mode decision procedure in H.264 video encoding. Then, the Lagrangian optimization approach is described in detail to find the optimal coding mode. Finally, to provide better understanding of the general coding procedure in H.264, the steps to encode a macroblock are explained.

4.1.2.1 Lagrange multiplier estimation for mode decision

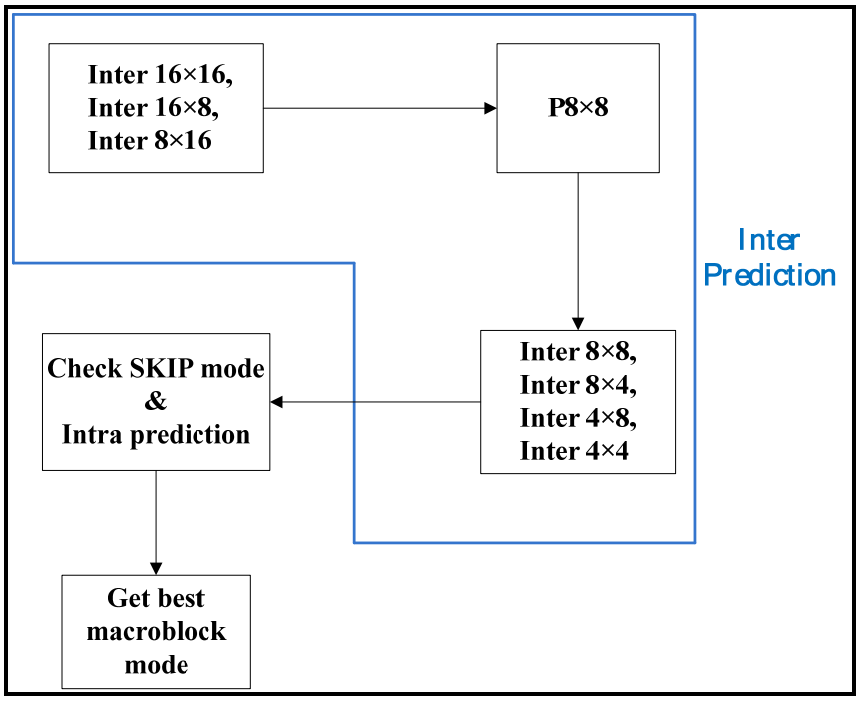

Figure 4.2 Macroblock mode decision program flow in a P-slice.

A key challenge in encoding H.264 is in mode selection. As discussed, H.264 has introduced many new features such as variable block sizes and various INTER, INTRA, and SKIP modes. So, to code a macroblock the encoder has to choose the best macroblock partition and

the mode of prediction for each macroblock partition from many different mode choices, such that the video coding performance is optimized. The selection process is conventionally referred to as macroblock coding mode selection (Xin, Vetro and Sun, 2004).

In figure 4.2, the mode decision program flow is shown when encoding a macroblock in a P-slice. For inter-coding P-pictures (or P-slices) in a progressive-scanned video, each 16×16 macroblock coding mode is varied over the following sets of possible MB modes (Lim, Sullivan and Wiegand, 2005):

- SKIP mode: the 16×16 mode where no motion and residual information is coded. In this instance, the motion vector predictor is applied to the region (macroblock).
- Four intra 16×16 modes.
- Nine intra 4×4 modes: each 4×4 partition can be coded using one of the nine prediction modes.
- 16×16 inter mode: coding of luma macroblock as one partition.
- 16×8 inter mode: coding the macroblock as two partitions. Each partition prediction can be done from different reference frame(s).
- 8×16 inter mode: coding of macroblock as two partitions with reference frame choices as above.
- 8×8 inter mode: coding of macroblock as four partitions. When 8×8 inter mode is considered, then each 8 × 8 sub-macroblock can further be independently partitioned into 8×8, 8×4, 4×8, and 4×4 blocks. The sub-macroblock partitions within an 8×8 sub-macroblock share the same reference picture(s).
    - 8×4 inter mode: coding of an 8×8 partition with two sub-partitions.
    - 4×8 inter mode: coding of an 8×8 partition with two sub-partitions.
    - 4×4 inter mode: coding of an 8×8 partition with four sub-partitions.

As well as the choice of prediction mode, the encoder has a wide choice of possible motion vectors. Each partition has its own unique motion vector for the use of motion-compensated prediction of the partition.

In an I-slice, the available modes for coding a macroblock are intra 4×4 prediction and intra 16×16 prediction for luma samples, and intra 8×8 prediction for chroma samples. As mentioned before, for 16×16 and 8×8 intra predictions, each 16×16 or 8×8 macroblock partition may be coded using one of the four defined prediction modes (Xin, Vetro and Sun, 2004).

By choosing each combination of coding modes, the encoder can generate a different distortion (reconstructed quality) and various numbers of coded bits ranging from low to high. Since every coding mode provides a different rate-distortion (RD) pair, the video encoder must choose the coding mode of a macroblock such that to achieve the best tradeoff between coded bitrate and decoded quality.

The target of rate-distortion optimization (RDO) is to minimize the overall perceived distortion D at a given rate budget R_c by an appropriate selection of the coding mode for each coding unit (Sullivan and Wiegand, 1998), namely

$$\min\{D\} \quad \text{subject to} \quad R \le R_c \tag{4.1}$$

If the distortion metric is additive, then D can be obtained by summing the distortion of all coding units, and the RDO problem can be formulated as follows in Eq. (4.2).

$$\min\left\{\sum_{i=1}^{N_u} d_i\right\} \quad \text{subject to} \quad \sum_{i=1}^{N_u} r_i \le R_c \tag{4.2}$$

where N_u denotes the number of coding units, d_i and r_i are the distortion and bitrate for the i^{th} coding unit, which may be a macroblock, a frame, or even a group of frames, and R_c is the maximum allowable bitrate.

There are two popular approaches to solve such a constrained problem: Lagrangian optimization and dynamic programming technique (Ortega and Ramchandran, 1998). In reality, the operational rate-distortion (RD) curve is composed of discrete operating RD

points. When the operating RD points on the RD curve are sparse and none of them meets the bit budget constraint, dynamic programming can still guarantee optimality, and unlike the Lagrangian technique can be able to reach points that do not reside on the convex hull of the RD characteristic (Ortega and Ramchandran, 1998). The computational complexity of dynamic programming is too high for practical applications, and is used only when direct Lagrangian optimization is difficult. Therefore, the Lagrange multiplier is widely used in today's video coding systems, including H.264, to solve the RDO problem (Richardson, 2010) due to its lower cost in computations (Lim, Sullivan and Wiegand, 2005), (Wiegand *et al.*, 2003a).

The basic idea of the Lagrangian technique is to convert the constrained optimization problem to an unconstrained one by the Lagrange multiplier method (Sullivan and Wiegand, 1998), which can be expressed as

$$\min\{J\} \quad \text{where} \quad J = \sum_{i=1}^{N_u} d_i + \lambda \sum_{i=1}^{N_u} r_i \tag{4.3}$$

where J is the RD cost function, and λ is known as the Lagrange multiplier which controls the tradeoff between distortion and bitrate. A smaller λ gives more emphasis to minimize distortion, generating a higher bitrate, whereas a larger λ will tend to minimize the rate at the expense of a higher distortion (lower decoded quality). For simplicity, it is assumed that the selection of the coding mode can be made independently for each coding unit without affecting the other units. Therefore, Eq. (4.3) can be reformulated as

$$\min\{J\} = \sum_{i=1}^{N_u} \min\left(d_i + \lambda r_i\right) \tag{4.4}$$

so that the minimum RD cost can be computed independently for each coding unit.

As discussed, the RDO typically uses a Lagrange multiplier to make the mode selection for each coding unit. A 16×16 pixel macroblock is the basic coding unit in the H.264/AVC. So,

the RDO evaluates the Lagrange cost J_{MODE} for each candidate coding-mode for the k^{th} macroblock S_k and selects the mode that gives the minimum cost. The Lagrangian cost function for the k^{th} macroblock is calculated as defined in Eq. (4.5).

$$J_{\mathrm{MODE}}\left(S_k, C_k, I_k \left| QP, \lambda_{\mathrm{MODE}}\right.\right) = D_{\mathrm{REC}}\left(S_k, C_k, I_k \left| QP\right.\right) + \lambda_{\mathrm{MODE}} R_{\mathrm{REC}}\left(S_k, C_k, I_k \left| QP\right.\right) \quad (4.5)$$

Where λ_{MODE} is the so-called Lagrange multiplier, D_{REC} represents the total distortion between the original macroblock S_k and reconstructed MB C_k samples, R_{REC} represents the rate after entropy coding (the number of bits required to encode the residue and motion vectors using mode I_k), and QP is the value of quantization parameter for transform coefficients. The MB coding mode I_k is varied over the sets of possible MB modes.

It is worth to note that for the SKIP mode, the distortion D_{REC} (D_{SKIP}) and the rate R_{REC} (R_{SKIP}) do not depend on the current quantizer value, and the distortion is measured between the original samples in the current frame and the inter predicted samples in the previous frame. The prediction samples are also selected from the reconstructed reference frame. The rate R_{SKIP} is the number of bits used to denote the skip macroblock type. Generally, R_{SKIP} has a very small value near zero and $D_{\mathrm{SKIP}} >> R_{\mathrm{SKIP}}$ (Ma, Gao and Zhao, 2009).

Figure 4.3 shows the process of computing the Lagrange cost for a coding mode of a macroblock partition. In this figure, the symbol T represents the H.264/AVC 4×4 transform. The difference between the input macroblock partition and its prediction is transformed, quantized, and then the rate is computed. To reconstruct the macroblock partition the inverse quantization and inverse transform are applied to the quantized transformed coefficients. After that, the distortion is computed between the reconstructed and the input macroblock partition. Finally, the Lagrange cost is computed using the computed rate and distortion.

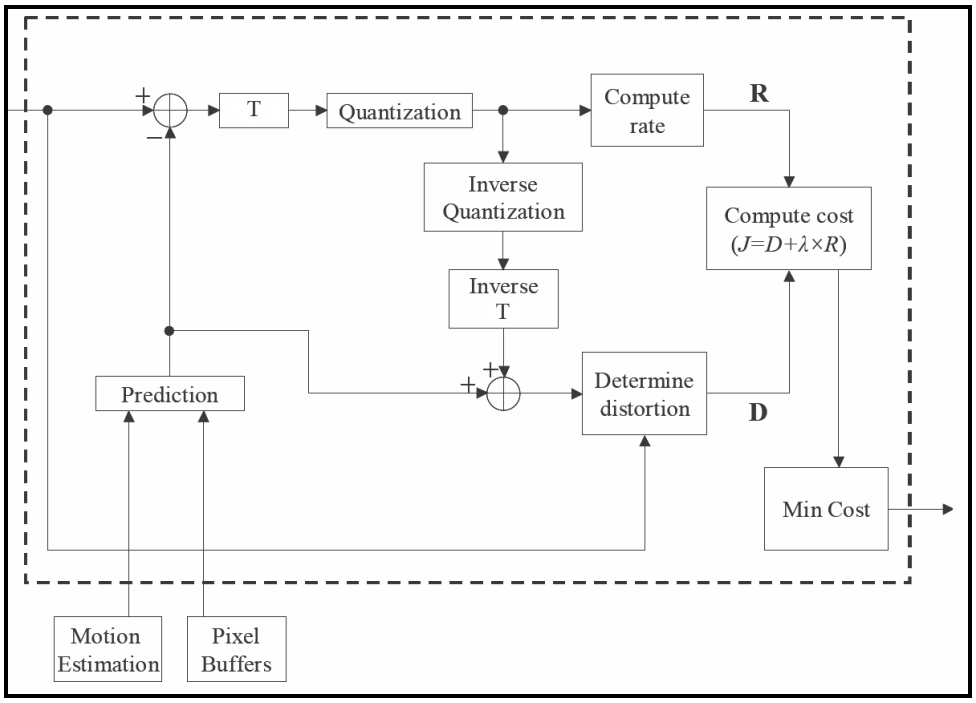

Figure 4.3 The RD cost computation process for a coding mode.
(Adapted from (Xin, Vetro and Sun, 2004))

The encoder typically selects the prediction mode for each block that minimizes the difference between a prediction block and the block to be encoded. The prediction residual is integrally transformed, quantized and transmitted using entropy coding together with the side information for indicating either intra or inter frame prediction. The best prediction modes are chosen by using the RD optimization which is described in Eq. (4.5). However, the distortion D_{REC} is calculated as the sum of squared differences or equivalently errors (SSD or SSE) between the original (S_k) and the reconstructed (C_k) macroblock pixels. The SSD (or SSE) is defined as:

$$D_{\mathrm{REC}}\left(S_k, C_k, I_k \middle| QP\right) = \sum_{(x,y)\in \boldsymbol{A}} \left|S_k(x,y) - C_k(x,y)\right|^2 \tag{4.6}$$

where (x,y) is the sample position in a subject macroblock $\boldsymbol{A}$. Other distortion metrics, such as sum of absolute differences (SAD) or sum of absolute Hadamard transformed differences (SATD) may be used in processes such as selecting the best motion vector for a block (Joint

Video Team (JVT) H.264/AVC Reference Software). Using a different distortion measure typically requires a different Lagrange multiplier calculation in the process.

As mentioned, to achieve the highest coding efficiency the encoder tries all the possible modes and chooses the best one in terms of least RD cost. The RD cost function in Eq. (4.5) is also used in the macroblock mode decision and the prediction mode decision of intra 4×4 and intra 8×8 modes. But the prediction mode decision for intra 16×16 does not involve RDO (Wiegand and Sullivan, 2003).

Assuming that the RD curve is convex, and both R_{REC} and D_{REC} are differentiable everywhere, the minimal J_{MODE} for a coding unit (macroblock) is given by setting its derivative to zero (Wiegand and Girod, 2001), i.e.,

$$\frac{\mathrm{d}J_{\mathrm{MODE}}}{\mathrm{d}R_{\mathrm{REC}}} = \frac{\mathrm{d}D_{\mathrm{REC}}}{\mathrm{d}R_{\mathrm{REC}}} + \lambda_{\mathrm{MODE}} = 0 \tag{4.7}$$

which yields

$$\lambda_{\mathrm{MODE}} = -\frac{\mathrm{d}D_{\mathrm{REC}}}{\mathrm{d}R_{\mathrm{REC}}} \tag{4.8}$$

Eq. (4.8) means that the Lagrange multiplier λ_{MODE} for the macroblock mode decision corresponds to the negative slope of the tangent to the RD curve of the prediction error coding (Wiegand and Girod, 2001). The Lagrange multiplier method itself does not point out how to determine λ_{MODE}. Calculating the optimal λ_{MODE} adaptive to video contents is a complex issue. To provide an effective choice of λ_{MODE} in a practical mode selection scenario, empirical approximations have been developed (Sullivan and Wiegand, 1998).

In general, there are two kinds of approaches to calculate the Lagrange multiplier. The first one is the heuristic way, where the Lagrange multiplier is determined by an iterative process or an empirical expression. There are different techniques in this approach such as a buffer

state based λ (Choi and Park, 1994) where the value of the Lagrange multiplier is a function of the current output buffer state by designing a feedback mechanism, and rate based λ (Wiegand *et al.*, 1996) which uses dynamic programming to find the minimal cost path in a trellis that each stage of the trellis corresponding to one macroblock and the branch cost is defined as the Lagrangian cost. In (Wiegand *et al.*, 1996), the Viterbi algorithm is used to obtain the least cost path through the trellis. The method in (Wiegand *et al.*, 1996) has been applied for H.263 video encoding, but employing such an algorithm for H.264/AVC with many coding parameter sets for a given coding unit would make the number of states too large for the computations to be tractable. In (Zhang *et al.*, 2007), (Zhang *et al.*, 2006), (Zhang *et al.*, 2010), the Lagrange multipliers are dynamically adapted according to the context (complexity) of the neighboring or upper layer blocks. In this context adaptive Lagrange multiplier method (CALM), the Lagrange costs of neighboring macroblocks are used to scale the Lagrange multiplier when calculating the cost of 16×16 blocks. Then the cost of 16×16 blocks can be used for determining the Lagrange multiplier for 16×8 and 8×16 blocks whose costs can in turn be used for estimating 8×8, 8×4 or 4×8 blocks, and so on.

It is worth mentioning that the performance of methods using heuristic approach is not generally satisfying due to the high computational complexity, and also there are no concrete theoretical foundations behind them. Therefore, a second approach is mostly used that determines the Lagrange multiplier in an analytical way (Wiegand *et al.*, 2003a), (Wiegand and Girod, 2001). Methods of this category have usually better computational efficiency and predictive accuracy. Analytical methods determine the Lagrange multiplier by using RD models and no iterative process is necessary. Thus, they can be very computationally efficient. Moreover, Eq. (4.8) implies that the more accurate the RD models are, the better λ_{MODE} can be achieved.

The H.264/AVC JM reference software uses an RD model under the high rate (HR) assumption to determine the Lagrange multiplier (Lim, Sullivan and Wiegand, 2005), (Wiegand *et al.*, 2003a). The RD function according to the typical high-rate approximation

curve for entropy-constrained scalar quantization is derived as follows (Wiegand and Girod, 2001)

$$R_{\mathrm{REC}}(D_{\mathrm{REC}}) = a \log_2 \left(\frac{b}{D_{\mathrm{REC}}} \right) \tag{4.9}$$

where a and b are parameterizing the relationship between rate and distortion, and their values depend on the video content. At sufficiently high rates, the source probability distribution can be approximated as uniform within each quantization interval (Gish and Pierce, 1968). So, the distortion model can be obtained as in Eq. (4.10).

$$D_{\mathrm{REC}} = \frac{Q_{step}^2}{12} = \frac{(2QP_{\mathrm{H.263}})^2}{12} = \frac{QP_{\mathrm{H.263}}^2}{3} \tag{4.10}$$

where Q_{step} is the quantization step size, and $QP_{\mathrm{H.263}}$ denotes the quantization parameter for H.263 coding standard. Note that the experiments in (Wiegand and Girod, 2001) that lead to the determination of the Lagrange multiplier have been conducted for the H.263 standard, and in H.263 the macroblock quantizer step size (Q_{step}) is twice the distance of the quantizer parameter ($QP_{\mathrm{H.263}}$).

Therefore, by taking the derivative of Eq. (4.9) and putting Eqs. (4.9) and (4.10) into Eq. (4.8), the final form of Lagrange multiplier is obtained.

$$\lambda_{\mathrm{MODE}} = c \cdot QP_{\mathrm{H.263}}^2 \tag{4.11}$$

where c is a constant. It is suggested by means of experimental results that 0.85 is a good value for the constant c (Wiegand *et al.*, 2003a), (Wiegand and Girod, 2001). The Lagrange multiplier for H.264/AVC can be obtained by considering the relationship between the quantization parameters in H.263 and H.264 standards, as defined in Eq. (4.12) (Wiegand)

$$QP_{\mathrm{H.263}} \approx 2^{(QP-12)/6} \tag{4.12}$$

where *QP* represents the macroblock quantization parameter in the H.264 coding standard. Combining Eqs. (4.11) and (4.12) yields the Lagrange multiplier for H.264 RDO mode selection process.

$$\lambda_{\text{MODE}} = 0.85 \cdot 2^{(QP-12)/3} \tag{4.13}$$

The Lagrange multiplier λ_{MODE} given by Eq. (4.13) is used for both Intra mode and P-slice Inter mode selection. But, for B slice testing the Lagrange multiplier $\lambda_{\text{MODE,B}}$ is determined as follows (Lim, Sullivan and Wiegand, 2005):

$$\lambda_{\text{MODE,B}} = \max\left(2, \min(4, \frac{QP-12}{6})\right) \times \lambda_{\text{MODE}} \tag{4.14}$$

In summary, this high rate λ selection method is practical and efficient. So, it has been adopted into the reference software of H.264/AVC (Lim, Sullivan and Wiegand, 2005), (Joint Video Team (JVT) H.264/AVC Reference Software). On the other hand, this algorithm also has some drawbacks. First, the Lagrange multiplier is only related to the macroblock quantization parameter and properties of the input video signal are not considered, that is, it cannot adapt itself to different video contents dynamically. Moreover, the high rate assumption of λ derivation is not always true, which will result in a poor performance in RDO process for low bit rate applications.

4.1.2.2 The conventional process of encoding a macroblock

In this subsection, we explain the complete process of encoding a macroblock when Lagrangian based RDO mode decision is in the high complexity mode. We mainly focus on the high complexity mode RDO since we perform our simulations in this mode.

Figure 4.4 depicts the block diagram of a standard video encoder including the motion estimation and mode decision blocks in the whole encoding process. Each frame is processed as a set of macroblocks, and each macroblock is subject to a transform, quantization, and

entropy coding. Motion estimation is carried out on the luminance signal, and a coding mode is selected considering the content of a pixel buffer. The error signal is produced by subtracting the input signal from the result of the prediction.

In the following, we explain the steps to encode one macroblock of a desired slice (I, P, or B slice) considering the mode decision procedure, when RDO performing in the high-complexity mode. We do not consider rate control in the encoding process in this subsection. Nevertheless, if rate control is enabled when encoding, the coding process of a macroblock related to the rate control is given by (Li *et al.*, 2003):

Rate control → Quantization parameter → RDO → MAD → Coding

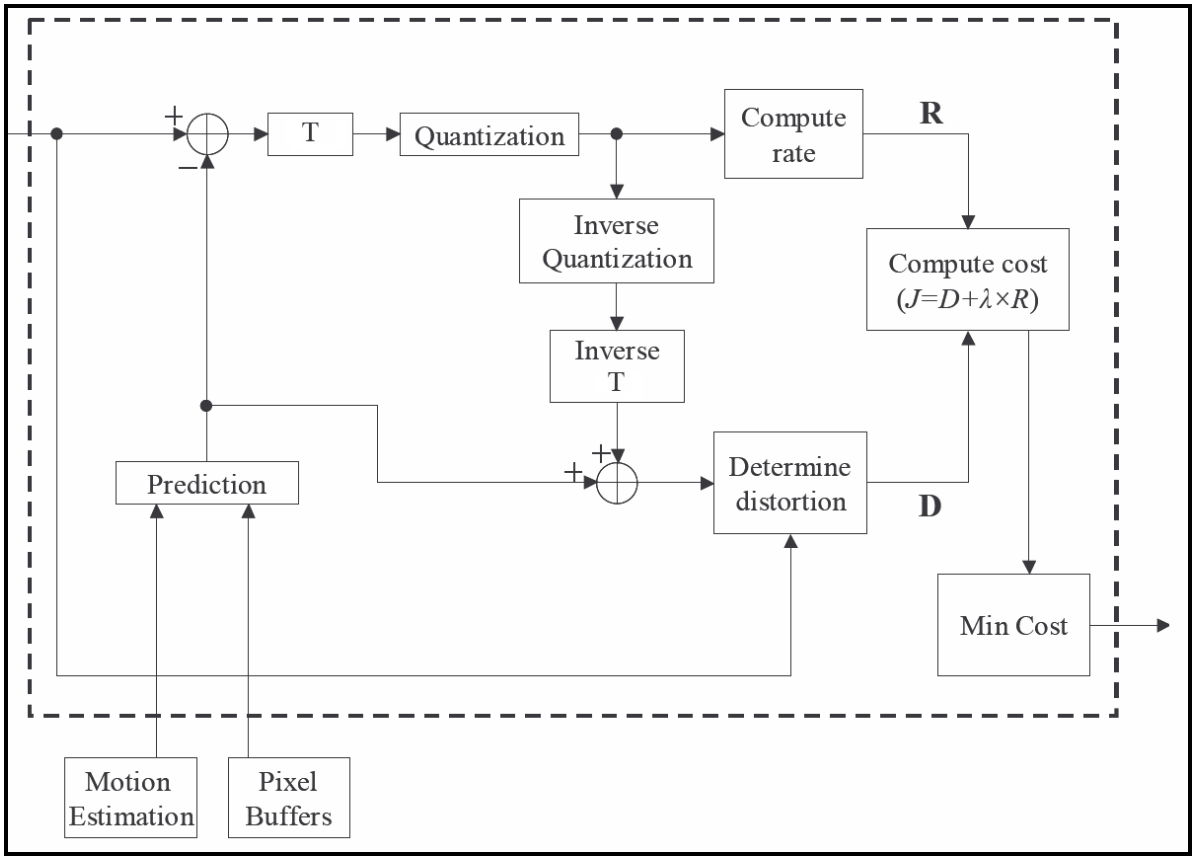

Figure 4.4 Block diagram of a hybrid video encoder including motion estimation and mode decision blocks.
(Adapted from (Xin, Vetro and Sun, 2004))

As mentioned earlier, the available macroblock coding modes in an I-slice include intra 4×4 and intra 16×16 for luma samples, and intra 8×8 for chroma samples. If a block or macroblock is encoded in intra mode, a prediction block is formed in the spatial domain based on neighboring encoded and reconstructed (but unfiltered) blocks of the same frame. This prediction block is then subtracted from the current block prior to encoding.

The inter mode decision process and coding in H.264 is much more computationally demanding than for intra modes, due to the block motion estimation step. For inter coding modes, a separate motion vector is required for each partition or sub-partition within the macroblock. In other words, the H.264 provides different motion compensated coding modes for macroblocks in P-slices.

Therefore, the inter mode decision process for the H.264/AVC encoding includes two steps. The first step is the rate-constrained motion estimation to search for the best matched blocks of the current encoding macroblock, from the reference frame(s) within a certain search range, for each inter prediction mode. In fact, the best matching block is chosen by using an RD function. The below Lagrange cost is used as the matching metric:

$$J(\mathbf{m}, \lambda_{\text{MOTION}}) = SA(T)D\left(S, C(\mathbf{m})\right) + \lambda_{\text{MOTION}} \cdot R_{\text{MOTION}}\left(\mathbf{m} - \mathbf{p}\right) \tag{4.15}$$

In the above formula, $\mathbf{m} = (m_x, m_y)^{\mathrm{T}}$ denotes the actual motion vector, and $\mathbf{p} = (p_x, p_y)^{\mathrm{T}}$ is the prediction for the motion vector. The R_{MOTION} is computed as the number of bits representing the predicted motion vector error information only and is computed by a table-lookup (Lim, Sullivan and Wiegand, 2005). λ_{MOTION} is the Lagrange multiplier for motion estimation. The metric $SA(T)D\left(S, C(\mathbf{m})\right)$ is the sum of absolute (transformed) differences between the original block *S*, and candidate matching block *C* (at the position designated by **m** in the reference picture). SAD is usually applied for integer pixel motion estimation while SATD is for subpixel (Wien, 2003). The block(s) with the minimum Lagrange cost *J* will be selected as the best matched block(s) for each prediction mode. In other words, for a block S_i, the Lagrangian cost function is minimized.

$$\mathbf{m}_i = \underset{\mathbf{m} \in \boldsymbol{R}}{\operatorname{argmin}} \{ J(\mathbf{m}, \lambda_{\text{MOTION}}) \} \tag{4.16}$$

with $\boldsymbol{R}$ being the search range, typically of ±32 integer pixels horizontally and vertically in H.264/AVC (Marpe, Wiegand and Gordon, 2005), and either one or more prior decoded picture is referenced. First, the motion search proceeds over integer-pixel locations to minimize Eq. (4.16). Then, half-pixel refinement is performed on the best of integer-pixel motion vectors to test whether a cost reduction in Eq. (4.16) is provided. Finally, the quarter-pixel motion search is applied to the previously determined half-pixel location to increase the accuracy of motion estimation. This subpixel refinement step yields the resulting motion vector $\mathbf{m}_i$.

The Lagrange multiplier λ_{MOTION} is used in computation of motion vectors in P or B slices. λ_{MOTION} is adjusted depending on the use of distortion measure in Eq. (4.15). In (Sullivan and Wiegand, 1998), (Wiegand and Girod, 2001) it is shown through experimental results that when the distortion in Eq. (4.15) is measured using *SAD* or SATD, λ_{MOTION} can be computed efficiently as

$$\lambda_{\text{MOTION}} = \sqrt{\lambda_{\text{MODE}}} \tag{4.17}$$

Correspondingly, when considering the SSD in Eq. (4.15), we would use

$$\lambda_{\text{MOTION}} = \lambda_{\text{MODE}} \tag{4.18}$$

The second step, after finding the best-matched blocks, in the H.264/AVC inter-mode decision process, is to select the best mode among the all candidate modes by computing the RD cost function, defined in the Eq. (4.5), for each prediction mode.

Generally, if there are N candidate modes for coding a macroblock, then the Lagrange cost of the n^{th} candidate mode J^n_{MODE}, can be computed as the sum of the Lagrange cost of its associated macroblock partitions (Xin, Vetro and Sun, 2004)

$$J_{\text{MODE}}^{n} = \sum_{i=1}^{P_n} J_{\text{MODE}}^{n,i} \qquad n = 1, 2, \ldots, N \tag{4.19}$$

where $J_{\text{MODE}}^{n,i}$ denotes the Lagrange cost of the n^{th} candidate mode for the i^{th} macroblock partition, and P_n is the number of macroblock partitions for the n^{th} candidate mode. The Eq. (4.19) is especially useful in intra mode decision, where the optimal coding-mode of each partition is chosen from several different candidate coding-modes.

When the i^{th} partition of the n^{th} macroblock has $K_{n,i}$ number of candidate coding-modes, the Lagrange cost of this macroblock partition is computed as defined in Eq. (4.20).

$$J_{\text{MODE}}^{n,i} = \min_{k=1,2,\ldots,K_{n,i}} \left(J_{\text{MODE}}^{n,i,k} \right) = \min_{k=1,2,\ldots,K_{n,i}} \left(D_{\text{REC}}^{n,i,k} + \lambda_{\text{MODE}} R_{\text{REC}}^{n,i,k} \right) \tag{4.20}$$

where $R_{\text{REC}}^{n,i,k}$ and $D_{\text{REC}}^{n,i,k}$ are respectively the rate and distortion of k^{th} candidate mode for the i^{th} partition of the n^{th} macroblock. After describing the details and steps of inter and intra mode selection, we can now summarize the procedure of encoding a macroblock, in the high-complexity mode, using the rate distortion optimization method as follows (Lim, Sullivan and Wiegand, 2005):

a) Perform motion estimation and reference frame selection for the modes inter 8×8, inter 8×4, inter 4×8, and inter 4×4 for each of the four 8×8 sub-macroblocks. The Lagrangian cost of the P8×8 mode is calculated by summing the cost for each sub-block together.
b) Perform motion estimation and reference frame selection for the modes inter 16×16, inter 16×8, and inter 8×16
c) Determine the best combination of intra modes
    - Select the best intra 4×4 prediction mode
    - Select the best intra 16×16 prediction mode
    - Compute the Lagrangian cost and find the best intra chroma mode
    - Compute the Lagrangian cost for the best intra modes

d) Choose the best prediction mode for the current macroblock using the results of parts a), b), and c)

It is particularly notable that when Eq. (4.5) is used for the final macroblock prediction mode decision, the distortion D_{REC} is considered as the overall distortion of luma (Y) component and chroma (U and V) components. For example, if SSD is used as the distortion metric, it is computed as follows in Eq. (4.21).

$$\begin{aligned} SSD(S,C,I \mid QP) = & \sum_{x=1,y=1}^{16,16} \left(S_{\mathrm{Y}}[x,y] - C_{\mathrm{Y}}[x,y,I \mid QP]\right)^2 \\ & + \sum_{x=1,y=1}^{8,8} \left(S_{\mathrm{U}}[x,y] - C_{\mathrm{U}}[x,y,I \mid QP]\right)^2 \\ & + \sum_{x=1,y=1}^{8,8} \left(S_{\mathrm{V}}[x,y] - C_{\mathrm{V}}[x,y,I \mid QP]\right)^2 \end{aligned} \tag{4.21}$$

The parameters $C_{\mathrm{Y}}[x,y,I \mid QP]$ and $S_{\mathrm{Y}}[x,y]$ represent the reconstructed and original luminance values; C_{U}, C_{V}, and S_{U}, S_{V} show the corresponding chrominance values.

Moreover, in practice R_{REC} in Eq. (4.5) is the number of bits associated with choosing mode and QP, including the bits for the macroblock header, the motion vector, the reference pictures, and the transformed Y, U, and V coefficients.

The Lagrangian based RDO procedure explained before, for encoding a macroblock, is used in case of high complexity mode decision. However, high complexity mode selection is computationally intensive, because for selecting the best macroblock mode it needs to compute the Lagrangian cost for all possible modes and full exhaustive search in motion estimation is used. So, a lot of real-time implementations and practical H.264/AVC encoders do not have the computational resources to carry out the full rate distortion optimized mode selection process in high complexity mode as described above. This practical constraint has led the development of various low complexity, and fast high complexity mode selection algorithms and approaches to speed up the original encoding process while maintaining the

quality of the reconstructed video. We give a brief review of different methods for fast high complexity mode RDO and low complexity mode decision in the annex II.

4.1.3 Adaptive and HVS-based Lagrange multiplier estimation in RDO for video coding

In this subsection, we first explain briefly how advanced and complex methods determine the Lagrange multiplier in the RDO process, and then describe the techniques in order to calculate the Lagrange multiplier so that the output visual quality conforms to the quality perceived by human observers.

4.1.3.1 Adaptive Lagrange multiplier calculation techniques

The aforementioned popular one-pass Lagrangian-based RDO algorithm, which is also recommended in the H.264/AVC reference software, performs the optimization only according to the quantization process while ignoring the properties of input video signals. In order to determine the Lagrange multiplier λ_{MODE} adaptive to video content, other methods have been proposed, like λ estimation method based on the ρ-domain technique, where ρ is defined as the percentage of zeros among the quantized transformed residuals (He and Mitra, 2002b), (He and Mitra, 2002a). In the ρ-domain method, the coding rate R and distortion D are considered as functions of ρ. It is shown that the rate function $R(\rho)$ is approximately linear given as

$$R(\rho) = \theta \cdot (1 - \rho) \qquad \text{bits per pixel} \tag{4.22}$$

where θ is a coding constant. Similarly, the distortion is defined as the mean square error introduced by quantization of the DCT coefficients. A simplified approximation to D is given as an exponential approximation in Eq. (4.23).

$$D(\rho) = \sigma^2 \cdot e^{-\alpha(1-\rho)} \tag{4.23}$$

The parameter α is a coding constant, and σ^2 denotes the variance of transformed residuals. Although the rate model is accurate, the distortion model is not always accurate for the purposes of rate control and thus, the frame bit allocation algorithm using the ρ-domain distortion model sometimes becomes unstable and fails to achieve the target bit rate (Kamaci and Altunbasak, 2004). Using the ρ-domain RD model and considering Eq. (4.8) at the same time, the Lagrange multiplier λ_ρ can be derived as shown in Eq. (4.24) (Chen and Garbacea, 2006).

$$\lambda_\rho = \beta \cdot \left(\ln\left(\frac{\sigma^2}{D} \right) + \delta \right) \cdot \frac{D}{R} \tag{4.24}$$

Both of parameters β and δ are coding constants. Due to the inclusion of the variance of transformed residuals into the Lagrange multiplier calculation, the Lagrange multiplier is able to adapt itself to input videos dynamically. However, the quantization parameter is not considered directly in the computation of λ_ρ, which makes the rate control difficult in some cases. Moreover, both R and D are directly included into the computation of the Lagrange multiplier, which may cause improper results because of error propagation.

To solve the above mentioned issues, a new Lagrangian RDO algorithm, namely Lap-λ, for one-pass coding is proposed in (Li *et al.*, 2007) and (Li *et al.*, 2009) based on the Laplace distribution of transformed residuals. The Lap-λ algorithm supposes that the transformed residuals follow a zero-mean Laplace distribution after motion estimation, and based on that the entropy of quantized transformed residuals is modeled for a uniform reconstruction quantizer. The entropy expression can be used to approximate the real rate. The rate R is obtained as shown in Eq. (4.25) (Li *et al.*, 2007).

$$\begin{aligned} R = & \frac{S \cdot e^{-\lambda_L (Q-F)}}{\ln 2} \left(-\left(1 - e^{-\lambda_L (Q-F)}\right) \ln\left(1 - e^{-\lambda_L (Q-F)}\right) \right) + \\ & \frac{S \cdot e^{-\lambda_L (Q-F)}}{\ln 2} \left(e^{-\lambda_L (Q-F)} \left(\ln 2 - \ln\left(1 - e^{-\lambda_L Q}\right) - \lambda_L F + \frac{\lambda_L Q}{1 - e^{-\lambda_L Q}} \right) \right) \end{aligned} \tag{4.25}$$

The distortion model is achieved as

$$D = \frac{e^{\lambda_L F}\left(2\lambda_L Q + \lambda_L^2 Q^2 - 2\lambda_L^2 QF\right) + 2 - 2e^{\lambda_L Q}}{\lambda_L^2\left(1 - e^{\lambda_L Q}\right)} \tag{4.26}$$

where S is a constant at sequence level to compensate errors resulted from non-ideal Laplace distribution, Q is the quantization interval (step), and F represents the rounding offset that equals to $Q/6$ for H.264/AVC inter-frame coding. λ_L is a distribution parameter called Laplace parameter and calculated according to the standard deviation of transformed residuals of each frame. λ_L is an inherent property of the input video, hence the proposed algorithm is able to achieve adaptivity according to the input sequences so that the overall coding efficiency is improved.

The proposed Lap-λ algorithm for RDO works on a frame level. It is supposed that the global optimality can be achieved when each frame is optimally coded. It seems nearly impossible to find an optimal Lagrange multiplier for a single macroblock by a statistical model. The reason is that in a small area, the related statistical properties, like standard deviation, are not representative, and thus the Laplace distribution may not be the best. However, considering adaptivity for bigger areas, such as foreground and background, may be an efficient approach.

The performance of the Lap-λ algorithm has been evaluated under common conditions for coding efficiency tests defined by JVT (Sullivan and Bjontegaard, 2001). In the environment of H.264/AVC baseline profile, significant gain of 1.79 dB in PSNR can be achieved for slow sequences by Lap-λ, but very limited gains are observed for complex videos with big movements, rotations or zooming, such as the “Mobile” and “Tempete” (Arizona State University Video Trace Library). The reason is that the residual variances for the fast video sequences are much bigger than those for slow ones because of less accurate predictions. Accordingly, the distribution of fast sequences is comparatively closer to the assumption of uniform distribution. Therefore, in such a case the high rate λ_{MODE} in Eq. (4.13) and the Lap-

λ share a similar performance, and they are close to each other. For slow sequences, the assumption of high rate λ_{MODE} fails, and λ_{Lap} is much bigger than the high rate λ_{MODE}, yielding a macroblock mode with less coding bits. In effect, the percentage of skipped macroblocks for slow videos is increased when the proposed Lap-λ is applied. Consequently, the bit-rate is greatly reduced at the cost of a bit higher distortion so that the overall performance is improved for the whole sequence (Li *et al.*, 2009).

In main profile, the overall performance is improved due to big λ_{Lap} applied on B-frames. Since the residuals in B-frames are generally quite small, λ_{Lap} becomes quite big so that the importance of lower rate is greatly emphasized in the RDO process. So, B-frames are coded with a little higher distortion but a much lower rate.

4.1.3.2 HVS-based Lagrange multiplier calculation techniques

The aforementioned content-adaptive Lagrange multipliers are not perceptual-based. The objective metric used for measuring the distortion in the RDO process has a strong impact on the quality of coded video. Widely adopted distortion metrics such as mean squared error, SSD or SAD, traditionally used in the RDO framework in H.264, are proved not correlating well with human perception (Wang and Bovik, 2009). To resolve this problem, some methods replace the SSD with HVS-related quality assessment metrics in the RDO framework of a video encoder. Among the various advanced image quality metrics, the structural similarity (SSIM) index has been more popular than others due to its prediction accuracy and computational efficiency. Therefore, some approaches have tried to consider and incorporate the SSIM as a quality metric into the RDO framework to improve the performance of a video encoder.

In (Yang *et al.*, 2009) an improved rate-distortion optimization method is proposed based on SSIM. The SSIM rather than SSD is adopted as the distortion metric in the RDO mode selection process. However, the SAD is still used in motion estimation to solve the problem of computational complexity. Therefore, the reconstructed macroblock obtained by each

inter-prediction mode remains the same of the conventional H.264/AVC encoding, and the rate-distortion gain is only obtained by better mode selection.

In order to choose the best prediction mode for each macroblock, the RD cost function for each prediction mode is defined as follows (Yang *et al.*, 2009)

$$J_{\text{MODE}}\left(S,C,MODE\middle|QP\right)=\lambda_{\text{MODE}}\left(1-SSIM\left(S,C\right)\right)+R\left(S,C,MODE\middle|QP\right) \quad (4.27)$$

where $SSIM(S,C)$ is the structural similarity between the original macroblock S and reconstructed macroblock C. Since the distortion calculated by $1-SSIM(S,C)$ is much smaller than the bit number calculated by $R(S,C,MODE|QP)$, the SSIM-rate multiplier λ_{MODE} is attached to the distortion term in the cost function. Basically, the Eq. (4.27) can be rewritten as

$$J=\lambda_{\text{MODE}}\cdot D_{FSSIM}+R \quad (4.28)$$

where D_{FSSIM} is the expectation of distortion $1-SSIM(S,C)$ and R represents the expectation of the bit number needed to encode one macroblock. The D_{FSSIM} is derived in the paper by performing experiments on a sequence with high-detail regions and high motion complexity (Mobile sequence) to include plenty of blocks types. The sequence is encoded by the H.264/AVC reference software and the average distortion of the reconstructed macroblocks is measured by (1-*SSIM*) for each frame in the sequence. The results of this experiment provide an approximate relationship between the macroblock quantizer value QP and the distortion D_{FSSIM}.

$$D_{FSSIM}=10^{-4}\cdot e^{\frac{QP+11.804}{6.8652}} \quad (4.29)$$

The encoded number of bits R, for a macroblock, is independent of the distortion metric used in the RDO process, and it is just related to the chosen prediction mode, quantization step and the matched macroblock. This means that, the R model remains the same, while SSIM is used

as distortion metric instead of SSD. Thus, by considering Eq. (4.8) the final λ_{MODE} is determined as

$$\lambda_{\text{MODE}} = 2.39 \cdot e^{\frac{QP+11.804}{6.8652}} \tag{4.30}$$

The simulations in (Yang *et al.*, 2009) demonstrate that the proposed technique has better RD performance, especially for middle-motion (or middle-complexity) video sequences or low encoding bit-rate, compared to conventional RDO in H.264/AVC. The reason is that for the low-motion (or low-complexity) sequences, most of the macroblocks are encoded by SKIP or 16×16 mode, and for the high-motion (or high-complexity) ones, most of the macroblocks are encoded by P8×8 mode, no matter which algorithm is used. Therefore, the encoding mode of a macroblock is relatively fixed in high or low-motion sequence coding. It is worth mentioning that for testing the proposed algorithm and comparing its performance with the original H.264 in inter coding, intra mode coding is forbidden in inter frame coding in both algorithms.

Authors in (Mai *et al.*, 2006) used the SSIM rather than the SAD as the distortion metric in the block matching motion estimation. The distortion using SSIM is measured as:

$$D(S,C) = 1 - SSIM(S,C) \tag{4.31}$$

where S and C are the original and the prediction block respectively. Therefore, the motion estimation Lagrangian cost function in Eq. (4.15) is rewritten as:

$$J(\mathbf{m}, \lambda_{\text{MOTION}}) = 1 - SSIM(S,C) + \lambda'_{\text{MOTION}} \cdot R_{\text{MOTION}}(\mathbf{m} - \mathbf{p}) \tag{4.32}$$

Due to the change of distortion measure, the Lagrange multiplier (λ'_{MOTION}) has to be modified correspondingly. The new Lagrange multiplier is determined from experiments, in conformity to the relation between $SSIM$ (S,C) and R_{MOTION}. In order to find the best matched block(s) and inter prediction mode for each macroblock, we calculate the total

Lagrange cost $J(\mathbf{m},\lambda_{\text{MOTION}})$ for each mode independently. The prediction mode with the minimum $J(\mathbf{m},\lambda_{\text{MOTION}})$ is chosen as the best inter prediction mode of the macroblock. The residual of this best mode is transformed, quantized and entropy coded. By using this approach, the RD cost function in Eq. (4.5) is not used, and consequently several macroblock coding processes are cut to reduce the computational load. It is shown that the SSIM-based motion estimation and mode selection method can reduce average 20% bit-rate and 2.5% of coding time while maintaining the same reconstructed video quality. The bit-rate reduction is due to the better matching function in Eq. (4.32) which uses SSIM, and the time saving is because of skipping RD based mode selection. It should be pointed out that the Lagrange multiplier has been determined only for $QP = 10$, and all the results generated based on that. Authors used full-search motion estimation method for performing the simulations and comparing with the conventional SSD technique.

In (Yang, Wang and Po, 2007), authors use similar SSIM-based approach to (Mai *et al.*, 2006) for motion estimation process, however, unlike the method in (Mai *et al.*, 2006), the RD cost function for mode(s) decision is defined based on SSIM metric. In the motion estimation part, the algorithm sets two fixed thresholds: a minimum motion vector (MV) cost and an early termination threshold. If the calculated cost of encoding current searching position's MV is larger than the minimum MV cost, it discards the current searching position, and goes to the next one. In RDO mode selection, the RD cost is calculated using the following equation

$$J_{\text{MODE}}\left(S,C,MODE\middle|QP\right)=K\left(1-SSIM\left(S,C\right)\right)+\lambda_{\text{MODE}}R\left(S,C,MODE\middle|QP\right) \qquad (4.33)$$

where K is a multiplier to enlarge (1-*SSIM*), obtained by experiments, and related to the *QP*. As explained before, the mode with minimum RD cost will be selected as the best prediction mode. The authors have tested their algorithm only for *QP*=10, 20, and 30, and the parameters used in the RD function are obtained experimentally. Their results show that the algorithm averagely improves compression ratio by 9.65% at the same coding time, and the

applied fixed threshold creates desirable results, especially for the sequences with little motion.

The SSIM-based mode selection techniques proposed in (Mai *et al.*, 2006) and (Yang, Wang and Po, 2007) do not specify a formula or an algorithm for calculating the Lagrange multiplier, and perform extensive experiments to find out the λ_{MODE} for each *QP*. In (Huang *et al.*, 2010), a perceptual approach is proposed in order to incorporate the SSIM as a quality metric into the RDO framework. A predictive method is developed to estimate the Lagrange multiplier, and applied to H.264 intra-frame and inter-frame mode decision. In the perceptual-based RDO, the SSIM index is used instead of the traditional SSD to measure the distortion between the original image block and the reconstructed image block. Thus, the cost function for H.264 mode decision is expressed in terms of the SSIM index.

$$J_{\mathrm{MODE}}\left(S,C,MODE\middle|QP\right)=D_{\mathrm{SSIM}}\left(S,C,MODE\middle|QP\right)+\lambda_{\mathrm{SSIM}}R\left(S,C,MODE\middle|QP\right) \quad (4.34)$$

The distortion metric D_{SSIM} is defined by

$$D_{\mathrm{SSIM}}\left(S,C,MODE\middle|QP\right)=1-SSIM\left(S,C\right) \quad (4.35)$$

In view of the fact that the prediction mode decision for I16MB mode does not involve RDO (Wiegand and Sullivan, 2003), this defined cost function is just used in the macroblock mode decision and the prediction mode decision of I4MB and I8MB modes in intra mode decision process. The SSIM index in Eq. (4.35) is calculated on a 4×4 block basis, and the D_{SSIM} of a macroblock is approximated by the sum of the D_{SSIM} of each individual 4×4 block. By using this approximation, the same Lagrange multiplier can be used for both prediction mode decision and macroblock mode decision.

It is empirically shown in (Huang *et al.*, 2010) that the tangent to the perceptual SSIM-based RD curve has very similar slope with the tangent to the MSE-based RD curve at the point that is closest to the tangent point on the perceptual SSIM-based RD curve. This is the

characteristic that is exploited to relate the Lagrange multiplier of the perceptual SSIM-based RDO to that of the MSE-based RDO. It is concluded from the experiments that the RD points obtained by the MSE-based RDO for different sequences can be approximately fitted with a power function of the form $D = \alpha R^{\beta}$. Therefore, it is possible to use the power function to approximate the MSE-based RD curve in the perceptual RD space. Figure 4.5 describes the general framework of encoding process using perceptual SSIM-based RDO.

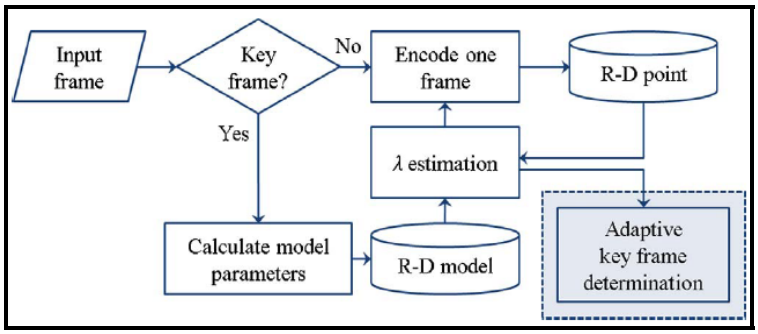

Figure 4.5 The general framework of video encoding using SSIM-based approach. (Adapted from (Huang *et al.*, 2010))

To estimate the Lagrange multiplier, we generate two distinct RD points by encoding a key frame using the MSE-based RDO for two different *QP* values. Then a curve fitting can be applied to these two calculated points to determine the two parameters α and β of the power function. Since consecutive frames in a sequence have high correlation with each other, and hence similar RD characteristic, the RD model of the key frame can be used as the predicted RD model of the subsequent frames. The Lagrange multiplier of a frame is determined by calculating the slope of the tangent to the predicted RD model, at the point that is closest to the RD point of the previous coded frame. In fact, the Lagrange multiplier value of each frame is obtained by using the R-D model of the previous coded key frame, and the RD point of the previous coded frame.

Two methods are described in (Huang *et al.*, 2010) to estimate the Lagrange multiplier : Gradient Descent and slope approximation approach. The slope approximation method is simple, and in opposition to gradient descent approach, no iteration is required. Simulation

results in (Huang *et al.*, 2010) show that the Lagrange multiplier values estimated by the gradient descent approach are very similar to that of the slope approximation approach.

For applying this perceptual SSIM-based method to mode decision of inter-frame coding, the key frame is adaptively determined. Let λ_t be the weighted average of the λs of the first five frames after the key frame. If the relative change of λ compared to λ_t is larger than a threshold, the next frame is set as a key frame (Huang *et al.*, 2010).

Briefly, the perceptual SSIM-based RDO uses the RD characteristics of a key frame to predict the RD characteristics of the subsequent frames till the next key frame appears. Simulation results show that, at the same SSIM value, the proposed approach achieves on the average 9% bit-rate reduction for intra-frame coding and 11% for inter-frame coding over the MSE-based RDO framework. Furthermore, the results indicate that the perceptual SSIM-based RDO has more performance gain at low bit-rates, and that gain is significant. It is also confirmed that under the high rate assumption, the Lagrange multiplier is simply just a function of *QP*, while in lower bit-rates it also depends on the content. The subjective test shows that the SSIM-based RDO preserves edge and avoids blocking artifact better than the MSE-based RDO. That is because the method takes the structural information into account, and avoids choosing prediction modes that introduce the structural distortion to the reconstructed image. The complexity overhead of the slope approximation is about 5%, almost all of which is due to the SSIM index computation.

Another SSIM-based Lagrangian RDO scheme has been proposed in (Wang, Ma and Gao, 2010), (Wang *et al.*, 2011), and (Wang *et al.*, 2012), which determines the Lagrange multiplier adaptively according to properties of input sequences. Similar to the approach in (Huang *et al.*, 2010), the SSIM is employed in the RDO process as the distortion metric, and the RD Lagrange cost can be calculated as defined in Eq. (4.34) and Eq. (4.35). To avoid discontinuities at the macroblock boundaries, the SSIM index is calculated with a larger window. For Y component, the SSIM index of the current macroblock to be encoded is calculated within a 22×22 block by a sliding window, but for Cb and Cr components 14×14

blocks are used. As explained before, the Lagrange multiplier is calculated by setting the derivation of RD cost to zero.

$$\lambda_{SSIM} = -\frac{\mathrm{d}SSIM}{\mathrm{d}R} = -\frac{\frac{\mathrm{d}SSIM}{\mathrm{d}Q}}{\frac{\mathrm{d}R}{\mathrm{d}Q}} \tag{4.36}$$

To derive the Lagrange multiplier adaptively, a statistical reduced-reference SSIM model and a source-side information combined rate model in the RDO process are used. The rate model is approximated based on an entropy model that excludes the bit rate of the skipped blocks (Wang *et al.*, 2012)

$$R = H \cdot e^{\zeta \Lambda Q + \psi} \tag{4.37}$$

where ζ and ψ are two parameters, and not very sensitive to the video content. Therefore, for both context-adaptive variable length coding (CAVLC) and context-adaptive binary arithmetic coding (CABAC) entropy coding methods, ζ and ψ are empirically set to be

$$\zeta = \begin{cases} 0.03 & B\ frame \\ 0.07 & otherwise \end{cases} \qquad \psi = \begin{cases} -0.07 & B\ frame \\ -0.1 & otherwise \end{cases} \tag{4.38}$$

The parameter Λ is supposed to be the Laplace parameter of the transformed residuals, and the entropy model H is adopted as defined in (Li *et al.*, 2009).

The SSIM model is estimated by a reduced-reference (RR) quality metric in the DCT domain which requires a set of features from the reference frame and quantization process for quality evaluation. To obtain the statistical properties of the reference signal and calculate the RR-SSIM metric, each frame is partitioned into 4×4 nonoverlapping blocks and DCT transform is performed on each block. Then, the DCT coefficients having the same frequency from each 4×4 window are grouped into one subband, which results in 16 subbands. The RR distortion measure is defined as

$$M_{RR} = \left(1 - \frac{D_0}{2\sigma_0^2 + C_1}\right)\left(1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{D_i}{2\sigma_i^2 + C_2}\right) \tag{4.39}$$

where N is the block size and σ_i is the standard deviation of the DCT coefficients from the i^{th} subband of the original frames. D_i represents the MSE between the original and distorted frames in the i^{th} subband, and is modeled by Laplace distribution of the residuals. Experimental investigations in (Wang *et al.*, 2011) and (Wang *et al.*, 2012) exhibit that the M_{RR} has a nearly perfect linear relationship with SSIM. Therefore, the RR-SSIM estimator $\hat{S}$ can be written as

$$\widehat{S} = \alpha + \beta \cdot M_{RR} \tag{4.40}$$

The parameters α and β are estimated, before coding the current frame, by using the straight line relating $\hat{S}$ and M_{RR}. Two points on the straight line are (1,1) and ($\hat{S}$, M_{RR}) from the previous frames of the same type. The experiments show superior performance of this SSIM-based scheme in rate reduction, while keeping the same level of SSIM quality value. For IPP GOP structure, 14% rate reduction is achieved on average in terms of SSIM index, and for the IBP GOP the rate reduction is averagely about 8%.

Calculating the Lagrange multiplier and mode selection in the aforementioned SSIM-based RDO frameworks are rather elusive. In (Yeo, Tan and Tan, 2012) the SSIM metric is described in terms of SSE and considered as the distortion metric in the RDO framework in an efficient manner by only scaling the SSE-based Lagrange multiplier without any changes to the RDO engine. The scale factor simply depends on the local variance in that region. Suppose x and y denote the original and the reconstructed image respectively. An additive distortion model is used for y, i.e. $y=x+e$, where e is the reconstruction error due to quantization. After simplifications, the distortion metric based on SSIM is defined as follows:

$$\mathrm{d}SSIM = \frac{1}{SSIM} \approx 1 + \frac{\mathrm{MSE}}{2\sigma_x^2 + c_2} \tag{4.41}$$

The assumptions involved in the defined distortion hold fairly well when the number of pixels in the region (N) is reasonably large. The RD cost for any block is defined by using d$SSIM$.

$$J_{\text{MODE}} = N \cdot \text{d}SSIM + \lambda R \tag{4.42}$$

After doing mathematical manipulations, the final Lagrange multiplier for the i^{th} macroblock is obtained as

$$\lambda_i = \frac{2\sigma_{x_i}^2 + c_2}{\exp\left(\frac{1}{M}\sum_{i=1}^{M}\log\left(2\sigma_{x_i}^2 + c_2\right)\right)}\lambda_{\text{MODE}} \tag{4.43}$$

where M is the number of macroblocks and λ_{MODE} is the SSE-based Lagrange multiplier defined in Eq. (4.13) previously. In fact, all that is required is to apply a local scaling of λ_{MODE} (the original Lagrange multiplier used in H.264/AVC JM reference software), depending on the local source variance and some source variance statistic computed over the entire frame. Therefore, the Eq. (4.43) provides a simple way of applying a small modification to the Lagrangian RDO mode selection process in order to maximize the SSIM over the entire frame. The implementation of the algorithm demonstrates that the proposed approach can achieve significant coding gains ranging from 3% to 18% for the same SSIM, comparing to original SSE-based implementation of JM software (Yeo, Tan and Tan, 2012).

It is worthy of mentioning here that some approaches, like the one proposed in (Wang, Li and Shang, 2007), have tried to consider the SSIM metric for perceptual image coding. The method in (Wang, Li and Shang, 2007) involves maximizing the minimal SSIM criterion using a bitplane based SPIHT image coder through an iterative optimization process. However, due to our focus on video coding, such perceptual coding methods are not explained in more details here.

There are another category of perceptual RDO mode selection methods which instead of directly optimizing the distortion metric, a locally varying perceptual-based Lagrange multiplier is used for RDO in each local region, alternatively (Yu *et al.*, 2005), (Sun *et al.*, 2007), (Lin and Zheng, 2008). This is a similar strategy to that proposed in (Yeo, Tan and Tan, 2012), however, the scaling and updating the Lagrange multiplier is done in a heuristic way, according to the perceptual characterises of video contents. In other words, the Lagrange multiplier in RDO process is adaptively updated according to the perceptual importance of each macroblock. The spatial and temporal characteristics of video contents may be utilized in perceptual importance analysis of macroblocks.

CHAPTER 5

THE PROPOSED PERCEPTUAL RDO BASED MODE DECISION USING LOW COMPLEXITY HVS RELATED DISTORTION METRICS

5.1 Motivation

In the previous chapter, we described the different approaches for Lagrangian RDO-based mode decision in H.264/AVC video coding. It was explained that the majority of available mode selection methods uses SSD (or equivalently SSE) as the distortion measure in computing the Lagrangian RD cost function. However, it is a widely accepted truth that the SSD has a poor correlation with human perception. Distorted frames with nearly equal SSD may have very different levels of perceptual distortion. This may cause the encoder, at a given rate, to generate a compressed stream which may not look as pleasing as it could to the human observer. To resolve this problem, various perceptual RDO mode selection techniques have been proposed.

Most of proposed HVS-based mode decision methods use the SSIM metric as the distortion measure in the RDO process. As mentioned earlier, the macroblock mode decision process is the second most computationally expensive phase in the encoding process. By considering the fact that the computational complexity of SSIM is much more than SSD, incorporating the SSIM into RDO process will add too much complexity to it, and makes the mode decision process too computationally burdensome.

On the other hand, the macroblock-based adaptive RDO schemes discussed earlier, like (Sun, Wang and Li, 2008) and (Chen and Guillemot, 2010), aiming to weight the Lagrange multiplier λ according to perceptual information and the properties of the HVS, still use the SSD as the distortion measure in the RDO process. Therefore, the decoded (or reconstructed) frame is finally optimized in terms of PSNR due to application of SSD as a distortion measure in the optimization objective function (Lagrangian RD cost function). Since, the

PSNR is not a reliable measure of perceived visual quality, the performance and subjective quality improvement of these λ-adaptive RDO methods are still in question.

Therefore, according to what explained above, an RDO mode selection method is desirable in video encoding that meets the following two criteria:

- The method should take into account the perceptual information in a sense that the final decoded frame is optimized in terms of an HVS-based quality assessment metric rather than PSNR.
- The method should be low in computational complexity, and not impose heavy computational costs to the RDO process.

Based on the above criteria, we incorporate our developed low-complexity quality metrics, described in the previous chapters, in the inter mode decision process of H264/AVC video encoding. To this end, we incorporate the $PSNR_A$ metric, or equivalently SSE_A, into the RDO process so that the amount of distortion is measured by SSE_A, instead of SSD, in the RD cost function. By this means, the final frame quality in the RD curve will be optimized in terms of $PSNR_A$, and as investigated before the $PSNR_A$ is a more accurate quality metric compared to the conventional PSNR. On the other hand, the computational complexity of $PSNR_A$ is very low, and hence, our method does not imply any extra computational load compared to the conventional RDO mode decision method in H.264/AVC.

In this chapter, we first formulate the theoretical framework of our proposed approach, and obtain the corresponding Lagrange multiplier for mode decision process. We explore how the Lagrange multiplier for our approach relates to the conventional mode decision SSD-based Lagrange multiplier in H.264/AVC video coding. We then apply an exhaustive search method to find out the optimal Lagrange multiplier for our method empirically. Finally, we investigate if our approach is really effective in improving the perceptual quality of decoded frames, and resulting in a better RD curve.

5.2 The proposed approach for perceptual coding mode decision

In the chapter 3, we explained that before calculating an image visual quality, it is better to apply wavelet transform to the image in order to increase the prediction accuracy. The required number of decomposition levels, *N*, for the Haar transform computed as

$$\mathrm{N} = \max\left(0, \mathrm{round}\left(\log_2\left(\frac{\min(H,W)}{(344/k)}\right)\right)\right) \tag{5.1}$$

where *H* and *W* are the height and width of the image respectively, and *k* denotes the viewing distance parameter. After applying the wavelet transform, the quality assessment metric $\mathrm{PSNR_A}$, or consistently the distortion metric $\mathrm{SSE_A}$, can be calculated using approximation subbands of the reference and distorted images.

$$\mathrm{SSE_A} = \mathrm{SSE}(\mathbf{X}_{\mathrm{A_N}}, \mathbf{Y}_{\mathrm{A_N}}) \tag{5.2}$$

To form our perceptual mode decision method, we use our metric, instead of SSD, to measure the distortion when calculating the RD cost function in the RDO process. Therefore, the distortion D_{REC} in Eq. (4.5) is substituted by our proposed distortion measure D_{P}.

$$D_{\mathrm{REC}}\left(S, C, MODE \middle| QP\right) = D_{\mathrm{P}}\left(S, C, MODE \middle| QP\right) = \mathrm{SSE_A} \tag{5.3}$$

Although the distortion metric defined as above in Eq. (5.3) for simplicity reason, it is also possible to use another form of $\mathrm{PSNR_A}$ for representation of the distortion measure. For example, the inverse of visual quality can be a measure of distortion degree. Therefore, another alternative is to define the distortion metric as the inverse of $\mathrm{PSNR_A}$ and hence, in this way the frames visual quality are maximized according to $\mathrm{PSNR_A}$, for a given rate.

$$D_{\mathrm{P}}\left(S, C, MODE \middle| QP\right) = \frac{1}{\mathrm{PSNR_A}} \tag{5.4}$$

Since the $PSNR_A$ and SSE_A are directly connected to each other by a bijective function (one to one mapping), the SSE_A is considered as the distortion metric for simplicity.

By definition of distortion metric, the mode decision RD cost function can be expressed as

$$J_{\mathrm{MODE}}\left(S,C,MODE\middle|QP,\lambda_{\mathrm{P}}\right)=D_{\mathrm{P}}\left(S,C,MODE\middle|QP\right)+\lambda_{\mathrm{P}}R_{\mathrm{p}}\left(S,C,MODE\middle|QP\right) \tag{5.5}$$

where λ_{p} denotes the Lagrange multiplier for the proposed approach. By employing Eq. (4.8) and supposing that R_{p} and D_{p} can be differentiable everywhere, the Lagrange multiplier λ_{p} is given by the following expression.

$$\lambda_{\mathrm{p}}=-\frac{\mathrm{d}D_{\mathrm{p}}}{\mathrm{d}R_{\mathrm{p}}}=-\frac{\dfrac{\partial D_{\mathrm{p}}}{\partial QP}}{\dfrac{\partial R_{\mathrm{p}}}{\partial QP}} \tag{5.6}$$

The encoded number of bits R_{p} for each macroblock is a function of its corresponding matched blocks, QP, and the selected prediction mode, that is, the bit rate R model is independent of the distortion metric used in the RDO mode decision. Therefore, the R_{p} model remains the same as R_{REC} given in the Eq. (4.9) as long as the motion estimation method is not altered. This implies that the $\lambda_{\mathrm{MOTION}}$ should be detached from λ_{MODE} and gets adjusted independently in computation of motion vectors in P or B slices.

In order to determine λ_{p}, the macroblock distortion $D_{\mathrm{p}}(QP)$ model is required for deriving the expression of $(\partial D_{\mathrm{p}} / \partial QP)$. It is very difficult to obtain $D_{\mathrm{p}}(QP)$ expression theoretically. As the SSE-based macroblock distortion model $D_{\mathrm{SSE}}(QP)$ is known to us in accordance with Eq. (4.10) and Eq. (4.12), the $(\partial D_{\mathrm{p}} / \partial QP)$ term is derived by using Eq. (5.7) if the macroblock distortion D_{p} (or $D_{\mathrm{SSE_A}}$) can be obtained in terms of D_{SSE} (or equivalently D_{SSD}).

$$\frac{\partial D_{\mathrm{p}}}{\partial QP}=\frac{\partial D_{\mathrm{SSE_A}}}{\partial QP}=\frac{\partial D_{\mathrm{SSE_A}}}{\partial D_{\mathrm{SSE}}}\cdot\frac{\partial D_{\mathrm{SSE}}}{\partial QP} \tag{5.7}$$

Therefore, determining the Lagrange multiplier λ_p theoretically is dependant to the problem of finding out a relationship between D_{SSE_A} and D_{SSE}.

In this chapter, we explore two approaches for establishing a relationship between macroblock distortion models in the pixel and wavelet domains: an analytical and empirical. Since there is no straightforward, direct relationship between D_{SSE_A} and D_{SSE}, we need to make some simple assumptions in our analytical method in order to be able to relate the two macroblock distortion models. Then, in the empirical analysis, we verify the validity of our assumptions and observe that if it would really be possible to get a reliable formulation between the two distortion models.

5.2.1 Theoretical analysis of macroblock distortions relationship

Let's suppose we replace SSE (or SSD) distortion measure by SSE_A in the RDO mode decision process. Now, our problem is to find a possible relationship between the two distortion measures. In addition, as we explained before, it is assumed that SSE_A is computed by using Haar wavelet transform. So, SSE_A-based block distortion is computed as:

$$\begin{aligned} D_{SSE_A} &= \sum_{(u,v)\in A_k^w} \left|S_w(u,v) - C_w(u,v)\right|^2 \\ &= \frac{1}{4}\sum_{(x,y)\in A_k} \left| \begin{matrix} \big(S(x,y)+S(x+1,y)+S(x,y+1)+S(x+1,y+1)\big) - \\ \big(C(x,y)+C(x+1,y)+C(x,y+1)+C(x+1,y+1)\big) \end{matrix} \right|^2 \\ &= \frac{1}{4}\sum_{(x,y)\in A_k} \left| \begin{matrix} \big(S(x,y)-C(x,y)\big)+\big(S(x+1,y)-C(x+1,y)\big)+ \\ \big(S(x,y+1)-C(x,y+1)\big)+\big(S(x+1,y+1)-C(x+1,y+1)\big) \end{matrix} \right|^2 \end{aligned} \tag{5.8}$$

where A_k and A_k^w represent the regions of a macroblock in the pixel domain and wavelet domain respectively, coded with mode I_K. For A_k region, we consider only those pixels with even positions in x and y. S and C are the original and reconstructed macroblock samples in the pixels domain, and S_w and C_w denote the original and reconstructed macroblock samples in the wavelet domain.

Now, for simplicity in notation, let the difference of two co-located pixels be denoted by d, as follows:

$$S(x,y)-C(x,y)=d(x,y) \tag{5.9}$$

So, Eq. (5.8) can be rewritten and expanded as

$$\begin{aligned}
D_{\mathrm{SSE_A}} &= \frac{1}{4}\sum_{(x,y)\in A_k}\left|d(x,y)+d(x+1,y)+d(x,y+1)+d(x+1,y+1)\right|^2 \\
&= \frac{1}{4}\sum_{(x,y)\in A_k}\begin{pmatrix} d^2(x,y)+d^2(x+1,y)+d^2(x,y+1)+d^2(x+1,y+1)+ \\ 2d(x,y)d(x+1,y)+2d(x,y)d(x,y+1)+2d(x,y)d(x+1,y+1)+ \\ 2d(x+1,y)d(x,y+1)+2d(x+1,y)d(x+1,y+1)+2d(x,y+1)d(x+1,y+1) \end{pmatrix} \\
&= \frac{1}{4}D_{\mathrm{SSE}}+ \\
&\quad \frac{1}{4}\sum_{(x,y)\in A_k}\begin{pmatrix} 2d(x,y)d(x+1,y)+2d(x,y)d(x,y+1)+2d(x,y)d(x+1,y+1)+ \\ 2d(x+1,y)d(x,y+1)+2d(x+1,y)d(x+1,y+1)+2d(x,y+1)d(x+1,y+1) \end{pmatrix}
\end{aligned} \tag{5.10}$$

Since, the last two terms of Eq. (5.10) are very complicated, it is not possible to directly relate or derive them in term of D_{SSE}. So, it is necessary to make a reasonable assumption in order to simplify them. Now, let's suppose the case where we have flat regions within 2×2 neighbourhoods of frames, such that the following relationship between pixels holds:

$$\begin{aligned}
&S(x,y)=S(x+1,y)=S(x,y+1)=S(x+1,y+1) \ \ \& \\
&C(x,y)=C(x+1,y)=C(x,y+1)=C(x+1,y+1)
\end{aligned} \tag{5.11}$$

Considering Eq. (5.9) and the above expression, yields the following equivalence:

$$d(x,y)=d(x+1,y)=d(x,y+1)=d(x+1,y+1) \tag{5.12}$$

Then, Eq. (5.10) is simplified to Eq. (5.13).

$$
\begin{aligned}
D_{\mathrm{SSE_A}} &= \frac{1}{4} D_{\mathrm{SSE}} + \\
&\quad \frac{1}{4} \sum_{(x,y)\in A_k} 3\left(d^2(x,y) + d^2(x+1,y) + d^2(x,y+1) + d^2(x+1,y+1)\right) \\
&= \frac{1}{4} D_{\mathrm{SSE}} + \frac{3}{4} D_{\mathrm{SSE}} = D_{\mathrm{SSE}}
\end{aligned}
\tag{5.13}
$$

It is observed that when Haar wavelet is applied for computing $\mathrm{SSE_A}$, the assumption of having flat regions causes the two macroblock distortion measures to be identical with each other, i.e. $D_{\mathrm{SSE}} = D_{\mathrm{SSE_A}}$. However, the estimation of flat pixel regions cannot always be realistic, especially for complex sequences. In order to verify the degree of validity of our assumption, we employ an empirical analysis of macroblock distortion measures. In our experimental analysis of distortion measures, we apply linear regression between SSE and $\mathrm{SSE_A}$ to check if there is any linear relationship between these two metrics. Before describing the experimental analysis of distortion measures, we explain about simulation conditions and details for evaluating or implementing our proposed algorithms in practice, for the rest of this chapter.

5.2.2 Simulation conditions

For our experimental analysis and evaluation of our proposed algorithm, the H.264/AVC reference software JM18.3 is adopted as our test platform. Although there are other implemented codecs of the H.264/AVC video compression standard, such as Intel IPP (OPAL Plug-In for Intel Integrated Performance Primitives, November 2012) and x264 encoder (x264: a Free H.264/MPEG-4 AVC Software Library and Application, 2012), we select the JM reference software in order to fully comply with JVT standard features.

In order to evaluate our proposed mode decision algorithm, we set common simulation conditions that follow the baseline profile. The baseline profile is specified in the H.264 standard document (Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264—ISO/IEC 14496-10 AVC), 2003). Our

simulations are done for the P-frames because B-frame coding is not included in the baseline profile. Test conditions and significant encoding parameter settings for the H.264 JM18.3 are tabulated in Table 5.1.

Table 5.1 Tabulation of significant encoding parameters for the H.264 JM18.3.

Parameter	Value	Description
GOP Structure	IPPP	P frames following an I frame
ProfileIDC	66	Baseline
LevelIDC	40	Supports 2Kx1K format. 62914560 samples/sec
Symbol Mode	0	CAVLC
NumberReferenceFrames	1	Number of previous frames used for inter motion search
FrameRate	30	Frame Rate per second
SourceWidth	352	Frame width
SourceHeight	288	Frame height
MDDistortion	2	Hadamard SAD as error metric for mode distortion operations
Transform8x8Mode	0	Only 4x4 transforms are used
ForceTrueRateRDO	1	No penalty for skip modes
RDOptimization	1	Enable high complexity mode
RateControlEnable	0	Disable rate control support
SearchMode	1	UMHexagon search
SearchRange	32	Allowable search range for motion estimation
FastCrIntraDecision	0	Disable performing a separate intra chroma mode decision

It is noteworthy that the JM software has certain limitations. In particular, its encoder is not able to perform all level/profile checks as specified in annex A of (ITU-T H.264 Telecommunication Standardization Sector of ITU, 2012). Therefore, the encoder may generate incompatible/non-conforming bitstreams which can not be decoded with other industrial decoders or video analysis software programs like Tektronix. In our case, the encoder limitations do not have any impact on the results and our conclusions. But when it is desired to analyze the encoded videos by other software tools, we can set the ProfileIDC =

100, LevelIDC = 22, and changing other encoding parameters within the JM configuration file (encoder.cfg) such that they obey the Baseline profile rules. The reason behind this setting is that the maximum number of motion vectors per two consecutive macroblocks can be 16 with the default settings (ProfileIDC = 66 and LevelIDC = 40). Therefore, a fully H.264 compatible decoder assumes that it is not possible to have all prediction modes at this profile (Baseline) and level.

In all of our simulations, we use video test sequences with CIF(4:2:0) YUV format. The test sequences are publicly available online through (Arizona State University Video Trace Library). All the test sequences are intra coded for the first frame (I-frame) and followed with subsequent inter coded frames (P-frames).

Finally, it should be noted that for performing our simulations and testing the algorithms, the JM software was run on a desktop PC with Ubuntu 10.04 (Lucid) LTS operating system, a 2.66-GHz Intel(R) Core(TM) 2 CPU, and 3 GB of RAM.

5.2.3 Empirical analysis of macroblock distortion measures

In order to explore the actual relationship between SSE and SSE_A, we perform a series of experimental simulations on different video sequences. The tests are conducted on three sequences with different motion characteristics, i.e. “container”, “foreman”, and “mobile”. Each sequence is firstly encoded with three different quantization parameters: QP = 16, 30, 44. The quantization parameters are ranging from low to high, which results in high to low bit-rates. The first 100 frames of each sequence are encoded for this test. Table 5.2 lists the bit-rates and the PSNR for Y component of each compressed test sequence at its corresponding QP.

In our verification process, the (SSE, SSE_A), (MSE, MSE_A), and (PSNR, $PSNR_A$) between the original macroblocks and their corresponding compressed macroblocks are calculated through sequence frames. Then, three statistical measures are used to assess the degree of

dependence between the computed pixel domain and wavelet domain distortion metrics in the previous step. The employed statistical measures are the Pearson correlation coefficient (LCC), Spearman rank correlation coefficient (SRCC), and Kendall rank correlation coefficient (KRCC).

Table 5.2 Comparison of the overall rates and distortions of H.264 compressed test sequences used in the experiment associated with investigating the relationship between distortion metrics SSE and SSE_A. The encoded test sequences are all in CIF resolution with the frame rate of 30 Hz.

Test Video	**QP = 16**		**QP = 30**		**QP = 44**	
	Y-PSNR (dB)	Bit rate (kbit/s)	Y-PSNR (dB)	Bit rate (kbit/s)	Y-PSNR (dB)	Bit rate (kbit/s)
foreman	45.095	3115.03	35.522	321.74	27.647	65.05
container	45.167	2002.05	34.751	145.33	26.811	17.41
mobile	44.463	7715.65	32.286	1502.60	22.075	122.40

To test linearity of the relationship between SSE and SSE_A, and between other metrics as well, a curve fitting is performed using linear regression between the two sets of metrics output data. The linear regression is done with the least squares method, and takes the following form

$$\begin{aligned} \mathrm{SSE}_\mathrm{A} &= \alpha \cdot \mathrm{SSE} + \beta \\ \mathrm{MSE}_\mathrm{A} &= \alpha \cdot \mathrm{MSE} + \beta \\ \mathrm{PSNR}_\mathrm{A} &= \alpha \cdot \mathrm{PSNR} + \beta \end{aligned} \tag{5.14}$$

where the two parameters α and β can be calculated simply, as explained in (Chong and Zak, 2011). The correlation coefficient LCC is calculated between the SSE scores and the objective model outputs after linear regression. Ideally, there should not exist any linear relationship between SSE and SSE_A. A linear relationship between SSE and SSE_A means that the Lagrange multiplier λ_p in Eq. (5.5) would be the same as the high rate λ_{MODE} defined in

Eq. (4.13) and hence, the selected macroblocks modes more or less remain the same as to those chosen by the traditional SSE-based method.

Tables 5.3-5.5 show the measures LCC, SRCC, and KRCC between pixel domain and wavelet domain macroblock distortion metrics in addition to regression parameters α and β for the sequence "foreman".

The results in Table 5.3 are for the case the distortion/quality metrics have been calculated for each 16×16 macroblock, and then metrics' values averaged for each frame over the whole sequence. But, the values in Table 5.4 and Table 5.5 are representative results of just a specific frame in the sequence.

As can be observed from the corresponding tables 5.3-5.5, the rank correlation coefficients (SRCC and KRCC) are increased with the increment of the *QP*. This means that when the perceptual video quality is high enough for the lower *QP* ranges, a low correlation exists between the two distortion measures and hence, it makes a difference which metric is used to measure the macroblock distortion. In contrast, for heavily compressed videos at low bit rates, there is a strong correlation between the SSE and SSE_A values and consequently it does not matter which metric is being used for macroblock distortion measurement. Our conclusion intuitively makes sense, since the perceptual quality is poor for heavily distorted videos. So, the difference between metrics judgements is negligible.

Moreover, it is found that the estimated regression parameters (α, β) vary with the quantization parameter value *QP*. Therefore, a fixed linear relationship does not exist between the two measures. This fact can also be confirmed by observing scatter plots in figures 5.1-5.6.

Table 5.3 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the sequence "foreman" (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock, and then metrics' values averaged for each frame over the whole sequence.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.9293	0.9099	0.9813
	SRCC	0.7827	0.8200	0.9529
	KRCC	0.5868	0.6509	0.8339
	α	0.2764	0.4661	0.7972
	β	7.6033	68.6995	-2.6123e+003
MSE vs MSE_A	LCC	0.9293	0.9099	0.9813
	SRCC	0.7827	0.8200	0.9529
	KRCC	0.5868	0.6509	0.8339
	α	1.1058	1.8642	3.1889
	β	0.1188	1.0734	-40.8172
PSNR vs $PSNR_A$	LCC	0.9211	0.9327	0.9862
	SRCC	0.8307	0.9355	0.9870
	KRCC	0.6400	0.7875	0.9067
	α	0.9157	0.8849	1.0487
	β	9.1339	7.1847	-0.0186

Table 5.4 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the frame number 25 of the sequence "foreman" (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.6890	0.8645	0.9535
	SRCC	0.5609	0.8729	0.9788
	KRCC	0.4020	0.6978	0.8785
	α	0.2087	0.3744	0.6275
	β	43.6130	519.6999	2.5505e+003
MSE vs MSE_A	LCC	0.6890	0.8645	0.9535
	SRCC	0.5609	0.8729	0.9788
	KRCC	0.4020	0.6978	0.8785
	α	0.8349	1.4975	2.5099
	β	0.6815	8.1203	39.8518
PSNR vs $PSNR_A$	LCC	0.7663	0.9484	0.9834
	SRCC	0.5607	0.8729	0.9788
	KRCC	0.4019	0.6979	0.8785
	α	0.6452	0.8923	0.9847
	β	21.3369	6.9510	1.8653

Table 5.5 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the frame number 75 of the sequence "foreman" (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.7608	0.8795	0.9548
	SRCC	0.6798	0.8892	0.9846
	KRCC	0.4993	0.7304	0.8964
	α	0.2008	0.3748	0.6190
	β	46.3166	594.3664	2.6438e+003
MSE vs MSE_A	LCC	0.7608	0.8795	0.9548
	SRCC	0.6798	0.8892	0.9846
	KRCC	0.4993	0.7304	0.8964
	α	0.8031	1.4994	2.4761
	β	0.7237	9.2870	41.3088
PSNR vs $PSNR_A$	LCC	0.8196	0.9645	0.9886
	SRCC	0.6798	0.8892	0.9846
	KRCC	0.4994	0.7304	0.8964
	α	0.5838	0.8294	0.9543
	β	24.1541	8.9381	2.6290

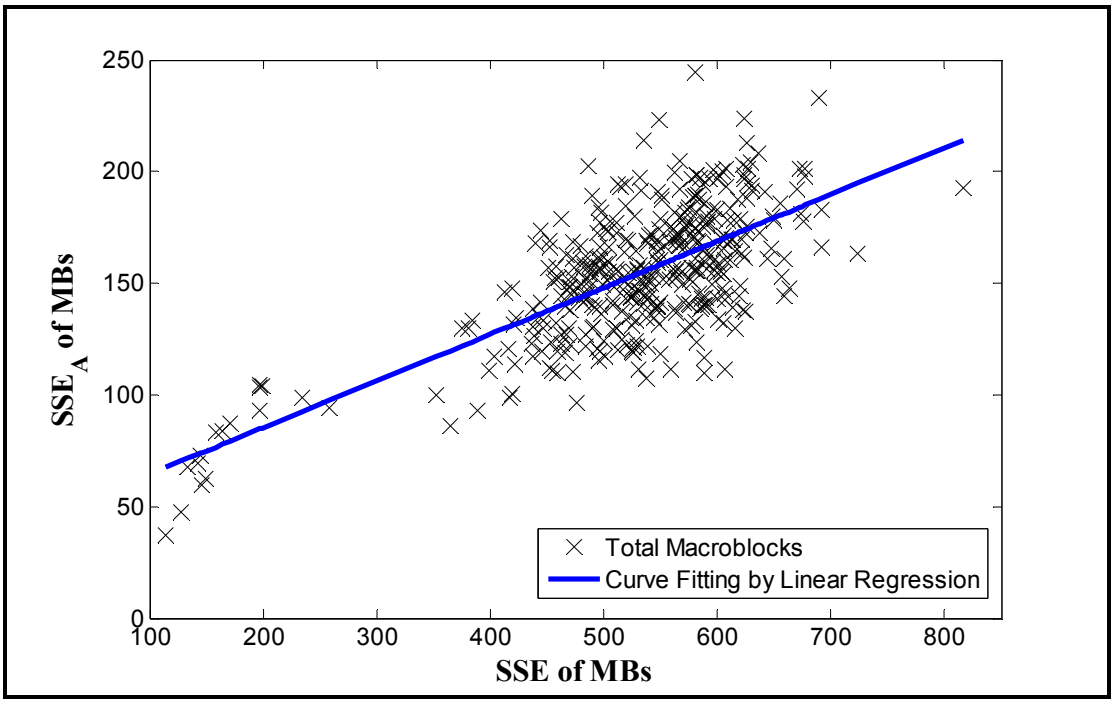

Figure 5.1 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "foreman" (CIF, 30Hz) coded with QP = 16. The distortion metrics calculated for each 16×16 macroblock.

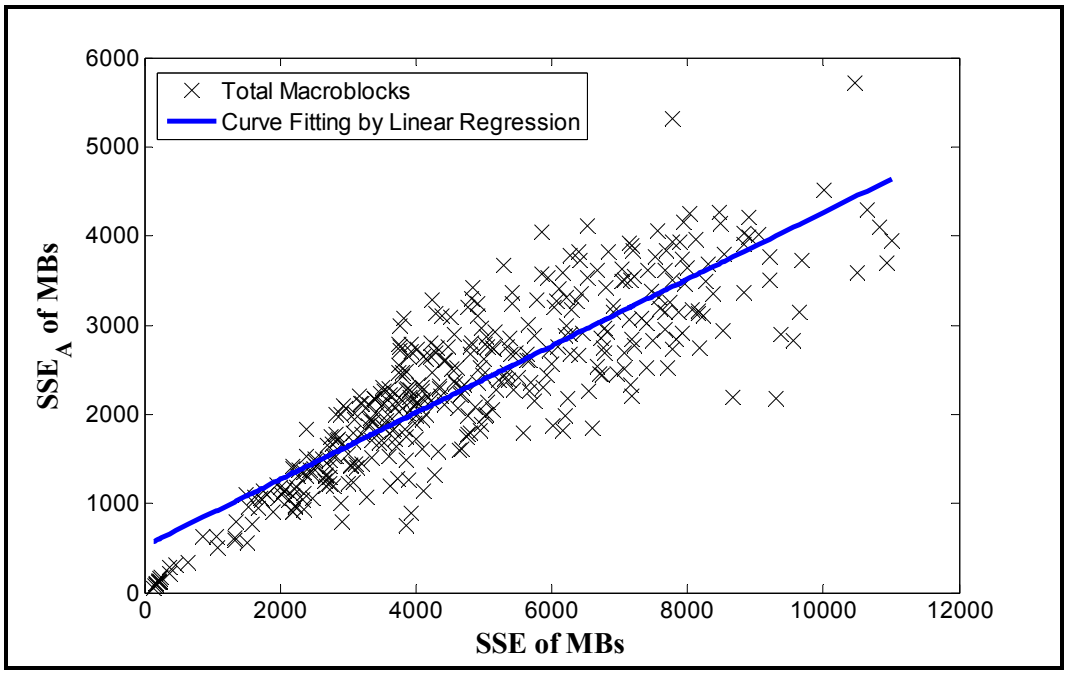

Figure 5.2 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "foreman" (CIF, 30Hz) coded with $QP = 30$. The distortion metrics calculated for each 16×16 macroblock.

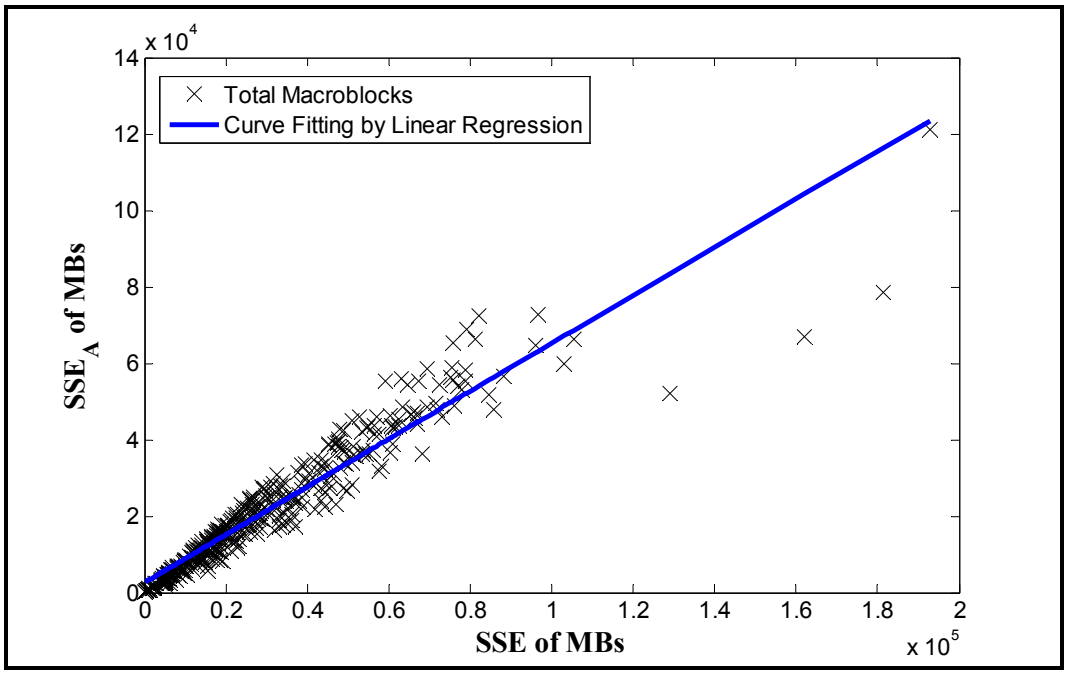

Figure 5.3 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "foreman" (CIF, 30Hz) coded with $QP = 44$. The distortion metrics calculated for each 16×16 macroblock.

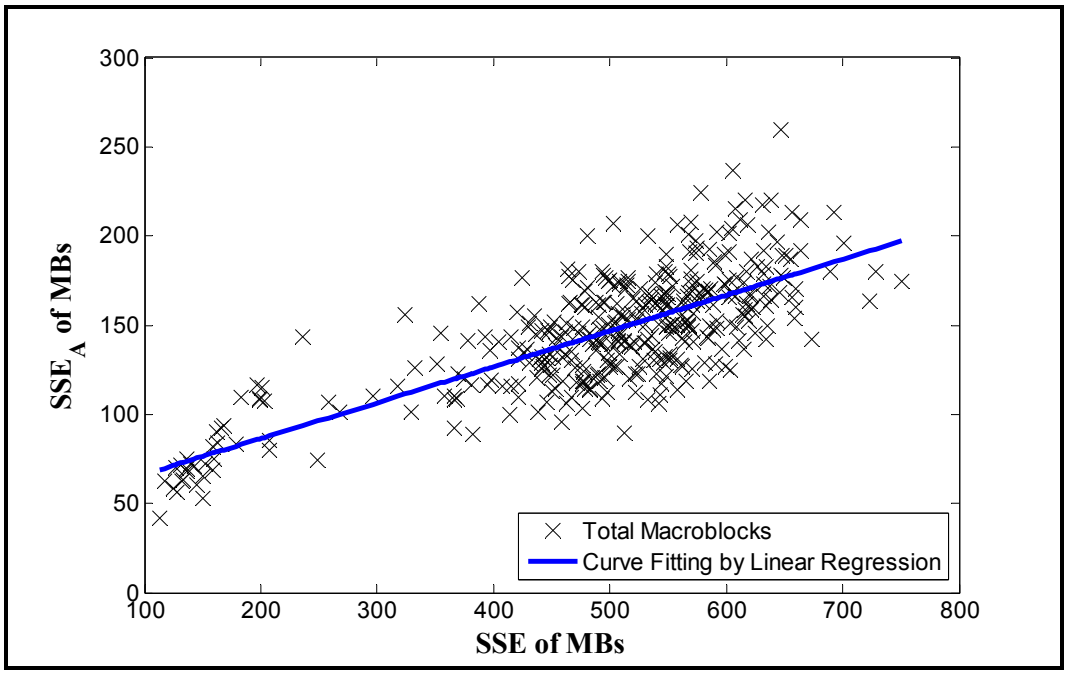

Figure 5.4 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "foreman" (CIF, 30Hz) coded with $QP = 16$. The distortion metrics calculated for each 16×16 macroblock.

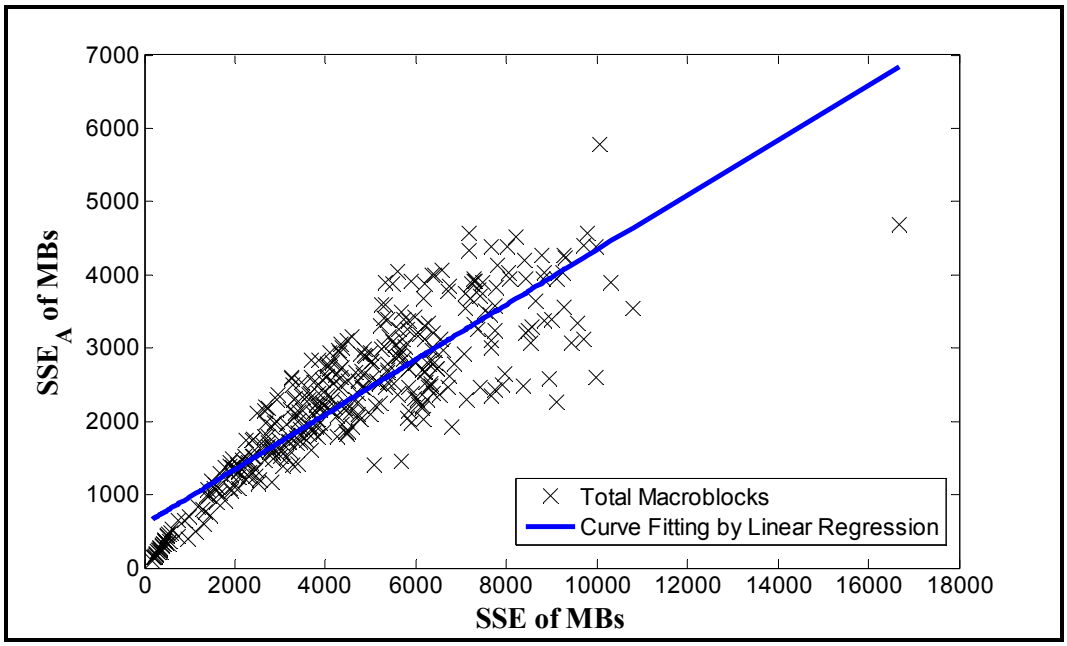

Figure 5.5 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "foreman" (CIF, 30Hz) coded with $QP = 30$. The distortion metrics calculated for each 16×16 macroblock.

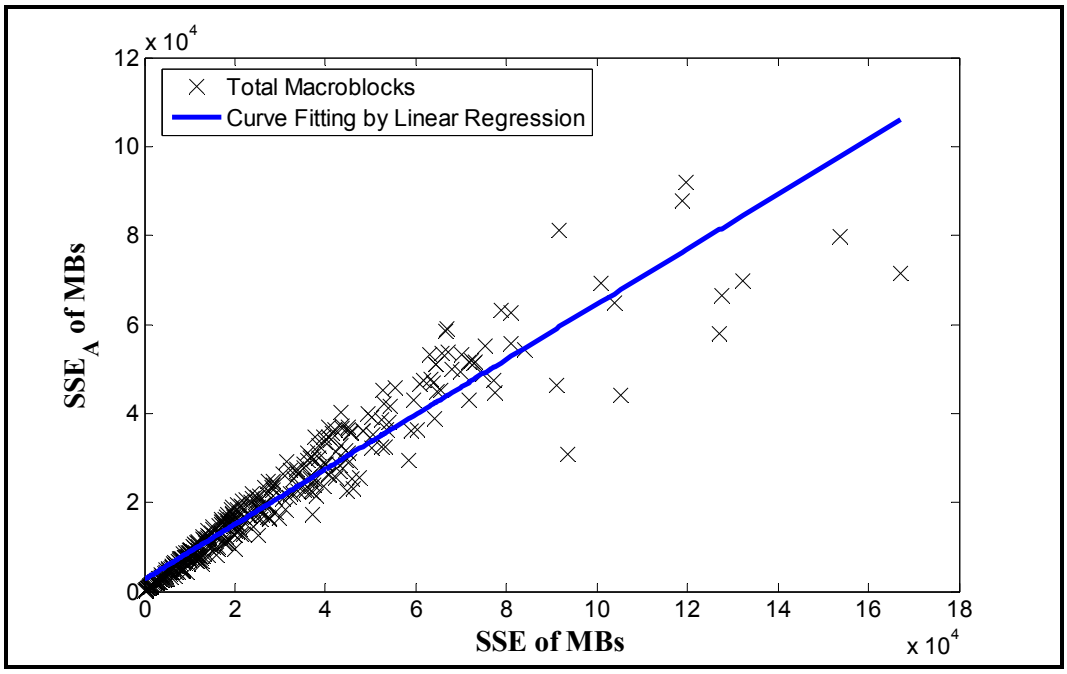

Figure 5.6 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "foreman" (CIF, 30Hz) coded with QP = 44. The distortion metrics calculated for each 16×16 macroblock.

Figures 5.1-5.6 show the scatter plots of SSE versus SSE_A macroblock distortion measurements for the frames 25 and 75 of the sequence "foreman" at three different QPs. Each sample point in the figures represents one macroblock distortion measurement. The fitted lines by Eq. (5.14) are shown in the figures as well. It is worth pointing out that the sample points in scatter plots of MSE versus MSE_A have shape and distributions exactly similar to those of shown for SSE and SSE_A. Therefore, we did not put scatter plots between MSE and MSE_A in the thesis to avoid repetitive results.

It is seen that at a low quantization parameter (QP = 16), the sample points are sparsely spread about the regression line in the figures 5.1 and 5.4. When the QP is increased, the sample points of scatter plots become more and more dense along the fitted line, and the SSE_A scores will be more consistent with the SSE ones. Furthermore, it can be observed that the sample points have more quadratic shape rather than a straight line. These observations confirm the results in tables 5.4 and 5.5, where LCC, SRCC, and KRCC values are increased as the QP increases.

All the results in the tables 5.3-5.5 and the figures 5.1-5.6 belong to the sequence “foreman”. The tables and scatter plots for the test sequence “mobile” have been brought in the annex III to keep the length of this chapter to a reasonable size. The results for other sequences are in the same order as those found for the sequence “foreman” and confirm the conclusions drawn based on tests on the sequence “foreman”.

Based on the above discussion, we conclude that using SSE_A as macroblock distortion measure in computation of RD cost function by Eq. (5.5) can be useful and produce some gain from high-rate to medium-rate coding. As it is demonstrated in table 5.6, the majority of macroblocks in a frame tend to be coded with SKIP mode at low bit rates. Thereupon employing a more accurate distortion measure for RDO mode decision would not be effective at low bit rate coding in order to gain a better rate-distortion performance.

Table 5.6 The frequency of each inter coding mode used, for macroblocks in the P slice, when encoding the first 100 frames of the sequence “foreman”.

Macroblock Mode	***QP* = 16**	***QP* = 30**	***QP* = 44**
0 (SKIP)	1194	14226	28268
1 (16×16)	8671	12059	7485
2 (16×8)	4581	4146	994
3 (8×16)	4701	4577	654
4 (P8×8)	18058	2751	67
5 (intra 4×4)	1493	733	113
7 (intra 16×16)	506	712	1623

Overall, we can conclude that there is not a simple relationship between the SSE and SSE_A measures especially at high bit rates and hence, finding out the Lagrange multiplier theoretically by using Eq. (5.7) would be very difficult. To resolve this problem, a brute-force search can be used to determine either the $\lambda_p(QP)$ or the $D_p(QP)$ function empirically. In the next section, we present our methodology to obtain the Lagrange multiplier λ_p, as a function of quantization parameter QP, for our proposed method.

5.3 The proposed search method for empirical determination of the Lagrange multiplier λ_p

As discussed before, when using the SSE_A as the distortion measure for the RDO mode decision in H.264 encoding, the conventional Lagrange multiplier in Eq. (4.13) can not work properly anymore, and hence it is necessary to find a new Lagrange multiplier for this new metric. In order to determine the Lagrange multiplier λ_p, we employ an exhaustive search method to find the best Lagrange multiplier value at each *QP*.

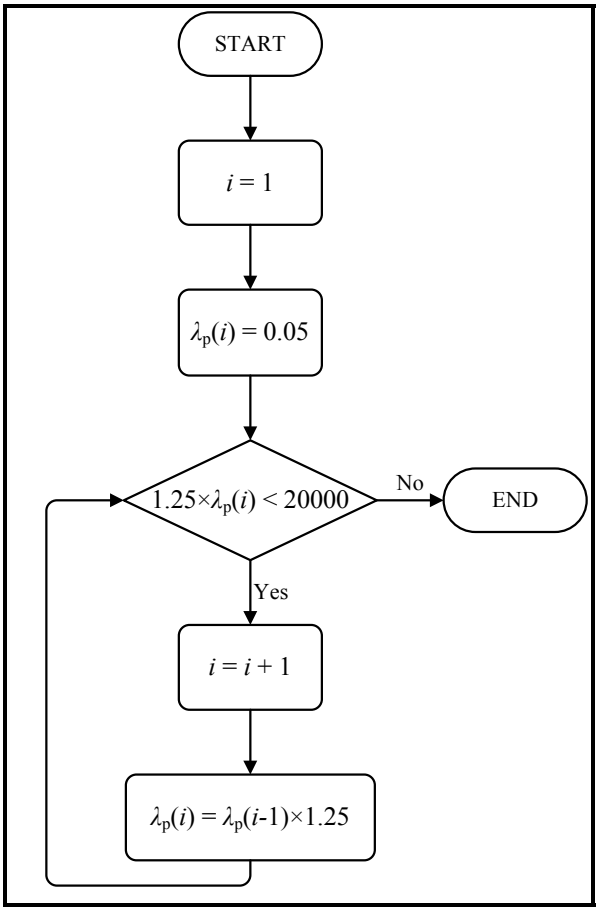

Figure 5.7 Diagram of Lagrange multiplier generation for our proposed search method.

In our search method, the effect of Lagrange multiplier setting is evaluated on overall rate-distortion performance of the video codec. Therefore, we search for the particular choice of the relationship between λ_p and *QP* that provides good results in terms of output rate-

distortion curve of the reconstructed videos. The bit rate is measured by kbit/s at frame rate of 30 Hz, and the distortion (quality) is represented by PSNR_A (dB) which is the average of all the encoded frames' PSNR_A. To accomplish our search, the H.264 coder is run using the proposed algorithm for the product space of different parameter sets of λ_p and *QP*. The *QP* ranges from 16 to 44 in steps of 2, i.e. $QP \in \{16, 18, 20,..., 44\}$, while the Lagrange multiplier varies over a selected range of values for each given QP. The diagram for the generation of the Lagrange multiplier set is shown in the figure 5.7. Using the given diagram, 58 values of Lagrange multiplier are generated which yields 870 combinations of the two parameters (*QP* and λ_p).

For each combination of the two parameters, three video sequences with different characteristics are encoded. The final rate-distortion curve for each sequence is formed by connecting the best rate-distortion points at each *QP*. By selecting the best RD point at a given *QP*, the associated Lagrange multiplier λ_p at that *QP* is found. An RD point is defined as a pair of rate–quality values $(R_{i,j}, \text{PSNR}_{\text{A}_{i,j}})$ computed at the QP_i and j^{th} Lagrange multiplier λ_{p_j}. The final rate-distortion curve of a test sequence is obtained by fitting an envelope to the fifteen different rate-distortion curves in which each belongs to a given *QP* for different λ_p values. For finding the final envelope and selection of best point on each RD curve, a main criterion is used: on the RD curve of a quantization parameter QP_i an RD point is considered invalid (non-optimal) point if the following conditions in Eq. (5.15) are met.

$$\text{if} \begin{cases} \left(\left(R_{i,j} \geq R_{i,k}\right) \& \left(\text{PSNR}_{\text{A}_{i,j}} < \text{PSNR}_{\text{A}_{i,k}}\right)\right) \;\; || \\ \left(\left(R_{i+1,j} \geq R_{i,k}\right) \& \left(\text{PSNR}_{\text{A}_{i+1,j}} < \text{PSNR}_{\text{A}_{i,k}}\right)\right) \end{cases} \Rightarrow \text{ invalid RD point} \qquad (5.15)$$
$$1 \leq i \leq 15 \;\; , \;\; 1 \leq j,k \leq 58$$

where $R_{i,j}$ and $R_{i+1,j}$ denote the bit rates of the encoded sequence for the Lagrange multiplier value λ_{p_j} at QP_i and QP_{i+1} respectively; $\text{PSNR}_{\text{A}_{i,j}}$ and $\text{PSNR}_{\text{A}_{i+1,j}}$ are the PSNR_A metric calculated between the original sequence and the compressed ones at rates $R_{i,j}$ and $R_{i+1,j}$. The

condition in Eq. (5.15) implies that the best points on two successive RD curves satisfy the following relation in Eq. (5.16).

$$QP_i > QP_{i-1} \quad \Rightarrow \quad \begin{cases} R_i < R_{i-1} \\ \mathrm{PSNR}_{\mathrm{A}_i} < \mathrm{PSNR}_{\mathrm{A}_{i-1}} \end{cases} \tag{5.16}$$

It should be mentioned that in our search algorithm, we start to evaluate the condition in Eq. (5.15) from the RD points generated with the lowest *QP* (16) to the highest *QP* (44), and then from high to low rates (low to high Lambda values). The RD points are evaluated sequentially, and each point is considered invalid if any of the two conditions is met for at least one *k*. We have demonstrated our MATLAB code, for fitting the envelope to different RD curves and finding the best point on each of the RD curves, in the annex IV to provide full guidance and give all details about it.

In order to verify the coding performance of our proposed algorithm and observe its amount of improvement, we compare the RD performance of our algorithm and that of the traditional MSE-based RDO method. Three video sequences from different classes of video test sequences are H.264 encoded with JM software for generating rate-distortion curves (Li *et al.*, 2001). Our selected test sequences have various spatial details and amount of movements. We chose sequences “container” from class A, “foreman” from class B, and “football” from class C (Li *et al.*, 2001).

In rate-distortion performance comparison between two mode decision algorithms, i.e. the proposed method and the traditional H.264 in inter coding, intra mode coding is disabled in inter frame coding in both algorithms. To disable all intra prediction modes for inter slices, the parameter “DisableIntraInInter” is set to 1 in the JM encoder software. All other parameters are set according to table 5.1. Since most of frames are inter-coded and also for simplicity, we implement and verify our algorithm just for inter modes, and intra mode decision is not modified from its traditional MSE-based version. Therefore, the Lagrange multiplier in Eq. (4.13) is still usable for intra mode prediction in the first I frame.

As our search algorithm for determining the Lagrange multiplier λ_p is not content adaptive, each sequence is encoded with two different numbers of frames, i.e. 30 and 120. By encoding 30 frames, the RD characteristic of the first frame well represents that of the subsequent frames for sequences with time-varying RD characteristic and low temporal correlation. But the results obtained for H.264 encoding of 30 frames and 120 frames would be very similar for slow sequences with small amount of movements like "container".

In the next section, we show the RD curves simulated by employing the explained search method for different cases of Lagrange multiplier.

5.4 Simulated rate-distortion curves

This section depicts the RD curves obtained by employing the search method explained before. We show the plots for different cases for which the curves have been simulated.

5.4.1 RD curves when $\lambda_{MOTION} = \lambda_{MODE}$

In this subsection, we bring the RD curves for the case when λ_{MOTION} is the same as λ_{MODE} (or λ_p). The λ_{MODE} values are assigned according to figure 5.7. As previously mentioned, the first frame is an I-frame which is encoded utilizing the traditional SSE-based mode prediction algorithm and Lagrange multiplier value, as used in the original JM software. Other P-frames are encoded using the described search method and the modified mode decision method.

Figure 5.8 demonstrates the concept of our search method. This figure shows 15 different rate-distortion (or rate-quality) curves for 15 values of QP and different values of λ_p. Each curve corresponds to a different value of *QP* and each marker on the curve represents a specific Lagrange multiplier. Figure 5.8 is for the "container" test sequence and 120 encoded frames.

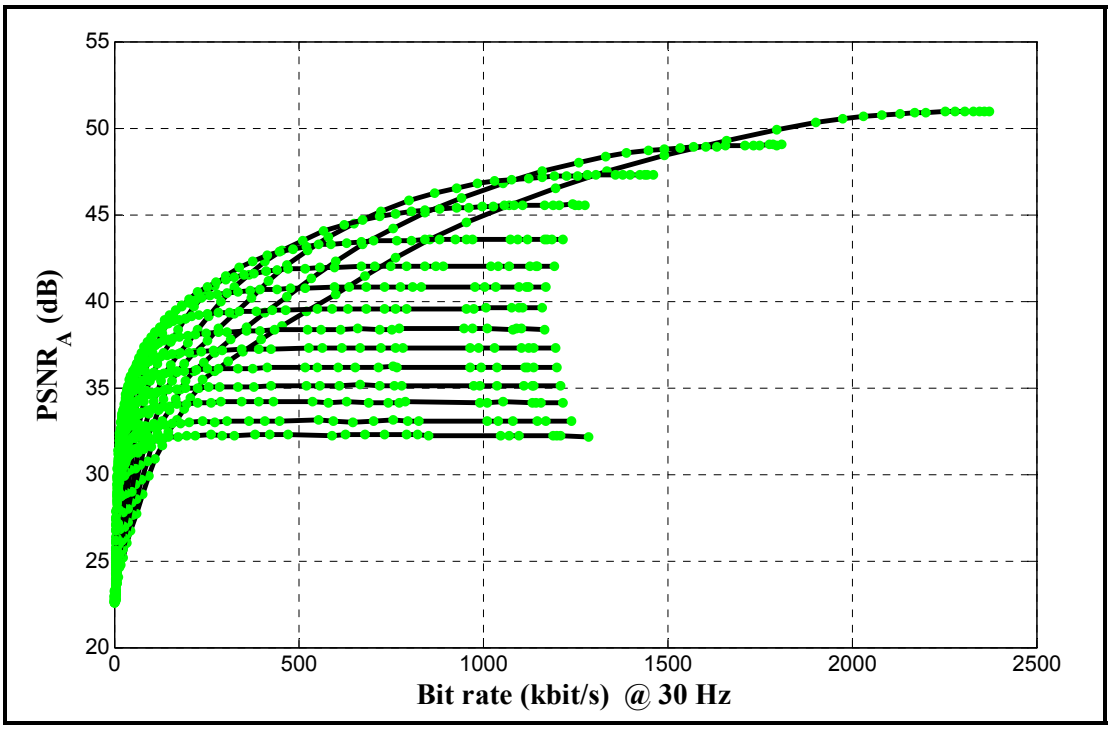

Figure 5.8 The 15 generated RD curves for 15 values of QP and 58 values of λ_p. The test sequence is "container" and the number of frames encoded is 120. Each curve corresponds to a different value of *QP* and each marker on the curve represents a specific Lagrange multiplier.

We have not showed the envelope of RD curves in figure 5.8 so that the curves to be more distinguishable. Figure 5.9 depicts the fitted envelope in addition to the generated RD curves in figure 5.8. The fitted curved has been represented by square Markers. From figure 5.9, it can be observed that the envelope curve obtained by our search procedure is tangent to each of the generated RD curves, and the search method selects the best point (Lagrange multiplier) on each curve very well.

To put our performance investigation in a proper context, we compare the RD performance of our algorithm with the traditional SSE-based mode decision algorithm that JM software applies in that process. Figures 5.10 to 5.12 show the rate-distortion curves for encoding 120 frames of the three test sequences. The dashed lines with left-pointing triangle markers belong to traditional SSE-based mode decision in the JM software, and the solid lines

represent the envelope of RD curves generated by SSE_A based mode decision. The generated RD curves for each *QP* are not shown in these figures to just focus on their final envelope.

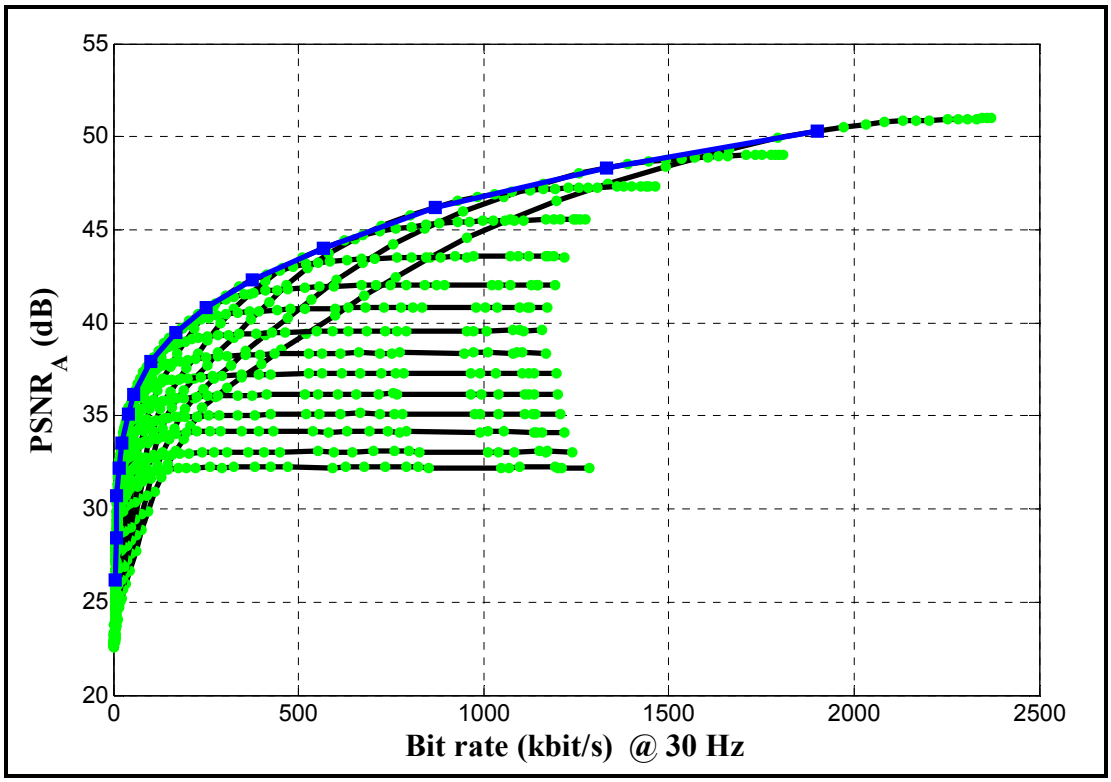

Figure 5.9 The generated RD curves (from figure 5.8) and the envelope fitted to them. Markers on the fitted curved have been represented by squares.

It is observed that the SSE_A based mode decision can achieve a better rate-distortion performance than the traditional SSE-based method. We can see that at the same $PSNR_A$ about 5% bit-rate reduction is gained by the SSE_A mode decision for the test sequences "container" and "foreman", however the performance gain is not noticeable for the sequence "football". This means that for sequences with big movements (low temporal correlation) most of the macroblocks are encoded by P8×8 mode; so, no matter which algorithm is used, the encoding mode of a macroblock is comparatively fixed.

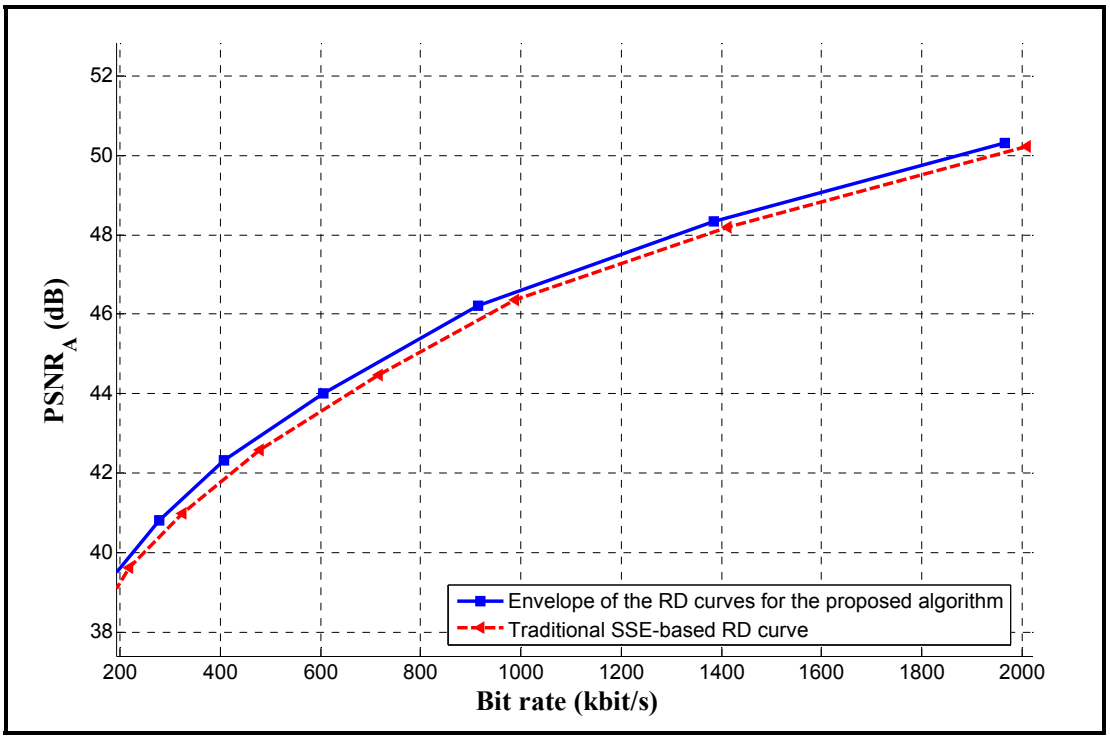

Figure 5.10 The rate-distortion curves for encoding 120 frames of sequence “container”.

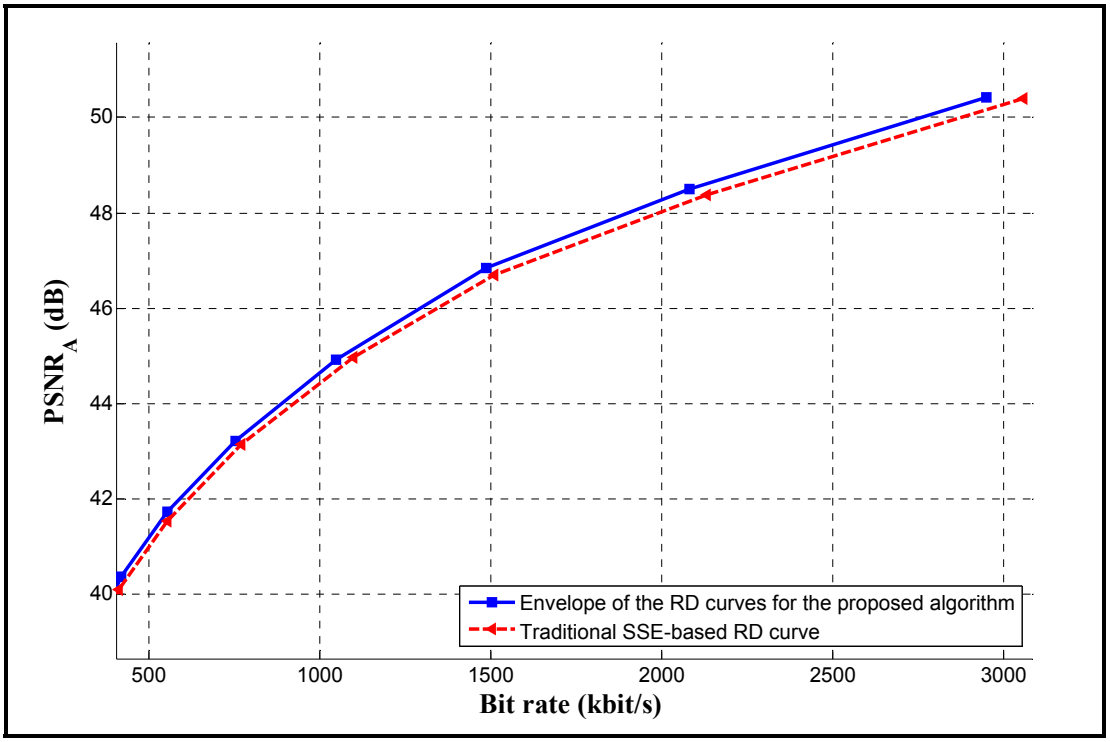

Figure 5.11 The rate-distortion curves for encoding 120 frames of sequence “foreman”.

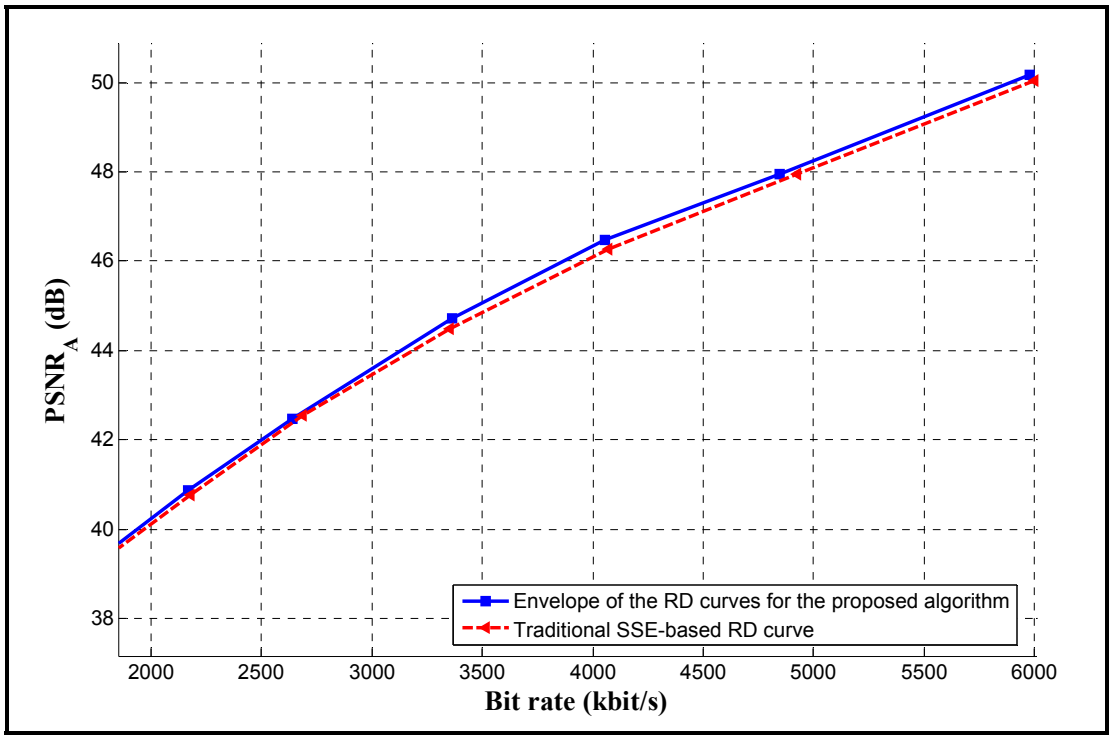

Figure 5.12 The rate-distortion curves for encoding 120 frames of sequence "football".

There are two causes that limit obtaining further RD performance gain for SSE_A-based mode decision. First, as discussed in section 5.3, our search method determines the Lagrange multiplier λ_p just as a function of *QP* and does not consider the content of frames in the video sequence. Therefore, with changes of the frames RD characteristics, the Lagrange multiplier cannot be adapted accordingly especially for fast sequences like "football". Second, the SAD is still reused in the motion estimation process of SSE_A-based mode decision and hence, the reconstructed macroblock by each inter-prediction mode would be the same as that of the conventional SSE-based encoding. Therefore, the RD performance gain is only due to better mode selection.

As we know, a frame RD characteristic does not usually vary substantially in a short period (like 30 frames) and subsequent frames maintain similar RD characteristics. Thus we repeat our simulations for the first 30 frames of each sequence in order to alleviate the former shortcoming. Figures 5.13 to 5.15 show the rate-distortion curves for encoding the first 30 frames of the three test sequences.

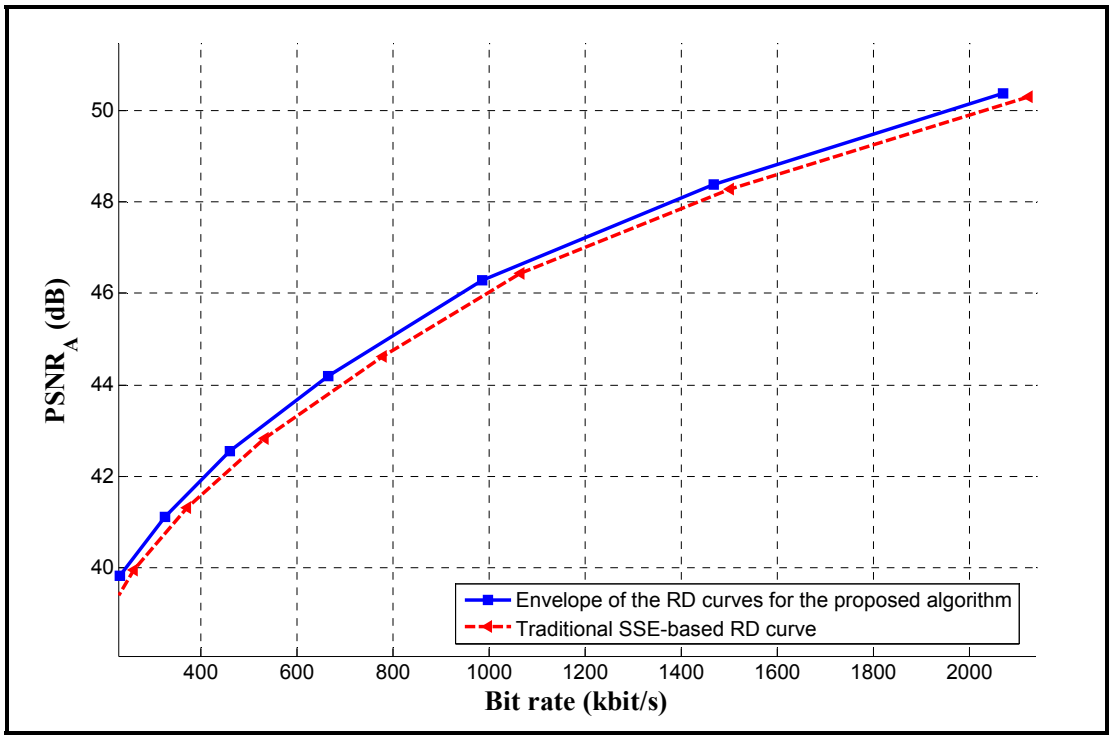

Figure 5.13 The rate-distortion curves for encoding 30 frames of sequence "container".

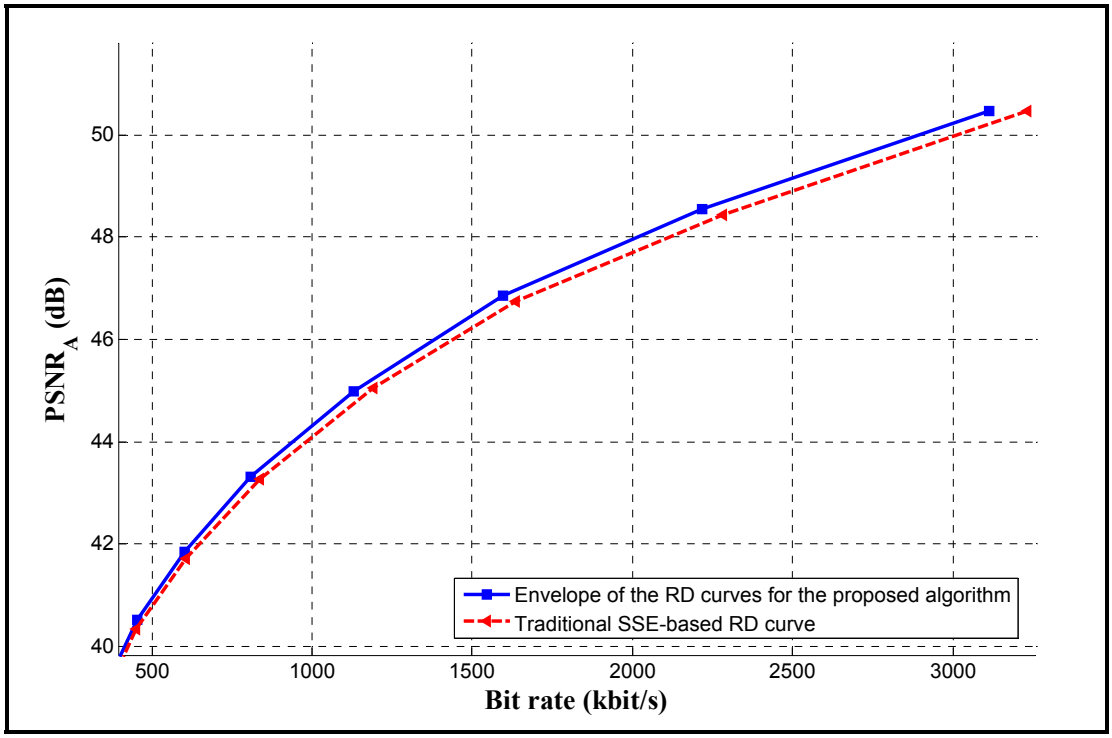

Figure 5.14 The rate-distortion curves for encoding 30 frames of sequence "foreman".

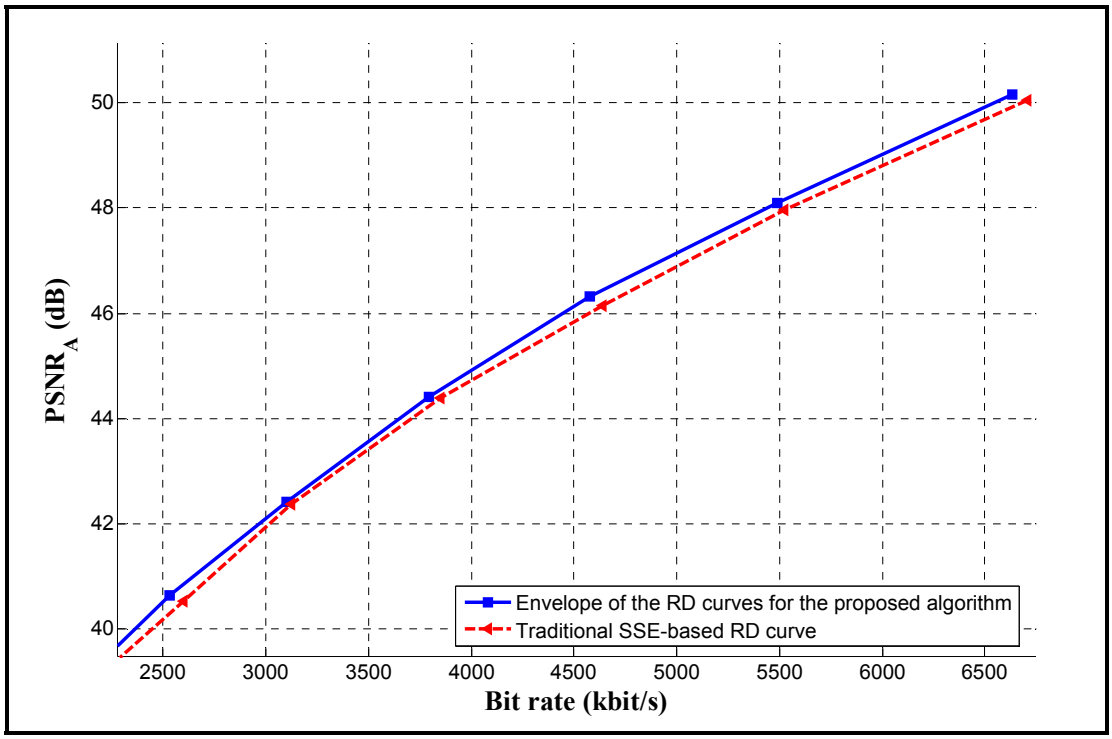

Figure 5.15 The rate-distortion curves for encoding 30 frames of sequence "football".

It is seen that the RD performance gain does not change remarkably for the sequences "container" and "foreman" compared with the case of encoding 120 frames. But a closer look at figure 5.15 indicates that RD gain improves for "football" in comparison with the case in figure 5.12. This gain improvement confirms the validity of our analysis about limitations of our search method, mentioned previously, for Lagrange multiplier determination.

Table 5.7 lists the adapted Lagrange multiplier (λ_p) values obtained by our search algorithm for encoding the three test sequences with two different numbers of frames, i.e. 30 and 120. The conventional SSE-based Lagrange multiplier λ_{MODE} (in Eq. (4.13)) values have also been shown in this table for comparison purpose. Figure 5.16 shows the adapted λ_p values for different *QP*s obtained by our search method. We observe that the adapted Lagrange multiplier varies with test sequences. This variation indicates that the rate-distortion tradeoff is dependant to the video content. It is apparent that the diversity of the adapted Lagrange multiplier λ_p for various sequences increases with *QP*. This observation is consistent with the

well-known fact that the Lagrange multiplier is simply a function of *QP* under the high rate assumption.

Moreover, it is seen that under the same *QP*, the fast/complex sequences with lots of details and big movements (like "football") use smaller λ_ps, because the quality of such sequences and their RD performance can be improved with a relatively small percentage of bit rate increase. For simple sequences, like "container", a higher percentage of bits can be saved at the expense of losing a relatively small amount of quality. As expected, we can notice that for a simple/slow sequence, e.g. "container", the Lagrange multiplier curves for a short period of 30 frames and the longer period of 120 frames are very similar and close to each other. That is because in a simple/slow sequence, the RD characteristics of consecutive frames vary relatively slow.

Table 5.7 Adapted mode decision Lagrange multiplier values obtained by the search method when $\lambda_{MOTION} = \lambda_{MODE}$.

QP	**Lagrange Multiplier λ_p**						**SSE-based λ_{MODE}**
	120 encoded frames			**30 encoded frames**			
	container	**foreman**	**football**	**Container**	**foreman**	**football**	
16	0.909	0.909	1.136	0.909	0.909	1.136	2.141
18	1.421	1.421	2.220	1.421	1.421	1.776	3.400
20	2.775	2.220	2.775	2.775	2.220	2.775	5.397
22	5.421	4.336	4.336	5.421	4.336	5.421	8.567
24	8.470	6.776	10.587	8.470	6.776	8.470	13.600
26	13.234	10.587	13.234	13.234	10.587	13.234	21.588
28	20.679	16.543	25.849	20.679	16.543	25.849	34.269
30	40.389	32.311	40.389	32.311	25.849	50.487	54.400
32	63.108	63.108	40.389	63.108	63.108	50.487	86.354
34	154.074	78.886	78.886	154.074	98.607	98.607	137.079
36	376.158	154.074	123.259	300.926	192.592	154.074	217.600
38	734.683	240.741	192.592	734.683	240.741	240.741	345.418
40	1147.943	470.197	240.741	918.354	470.197	376.158	548.317
42	2242.077	734.683	376.158	1793.662	918.354	587.747	870.400
44	4379.057	1147.94	734.683	3503.246	1793.66	1147.94	1381.673

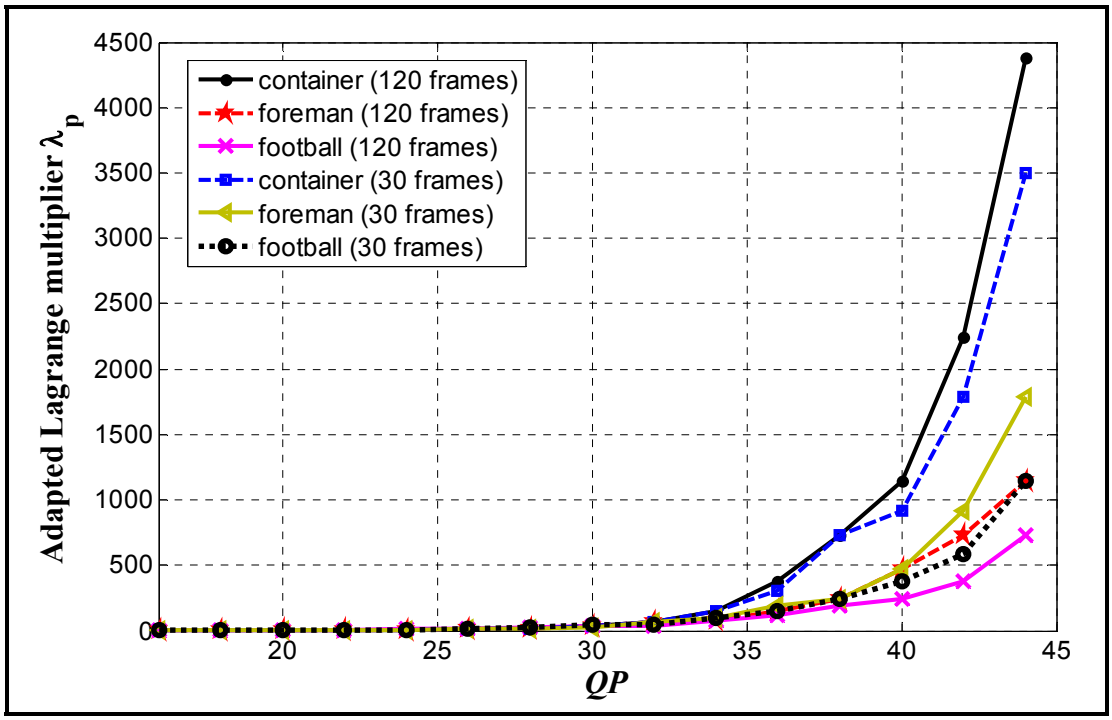

Figure 5.16 Adapted Lagrange multiplier values for various test sequences.

5.4.2 RD curves when $\lambda_{MOTION} \neq \lambda_{MODE}$

As mentioned before, our SSE_A-based approach for mode decision is still using SAD for motion estimation (at Full-Pel layer). Therefore, we may expect that the Lagrange multiplier for motion estimation λ_{MOTION} doesn't need to be changed from its conventional form in the original JM software. To investigate the impact of λ_{MOTION} on SSE_A-based mode decision, we repeat our tests in the previous subsection supposing that λ_{MOTION} is disconnected from λ_{MODE}, and takes its values according to the formula brought in Eq. (4.13). The λ_{MODE} will still be changed and set according to what explained in our search method, and its values are generated based on the diagram in figure 5.7. Figures 5.17-5.22 show the RD curves of SSE-based and SSE_A-based mode decision for various test sequences when $\lambda_{MOTION} \neq \lambda_{MODE}$. It can be observed that our assumption about λ_{MOTION} does not create a noticeable improvement on the RD performance of H.264 coding, relative to the previous case of $\lambda_{MOTION} = \lambda_{MODE}$. This observation is an indication that the H.264 coding performance is not very sensitive to variation of the λ_{MOTION}.

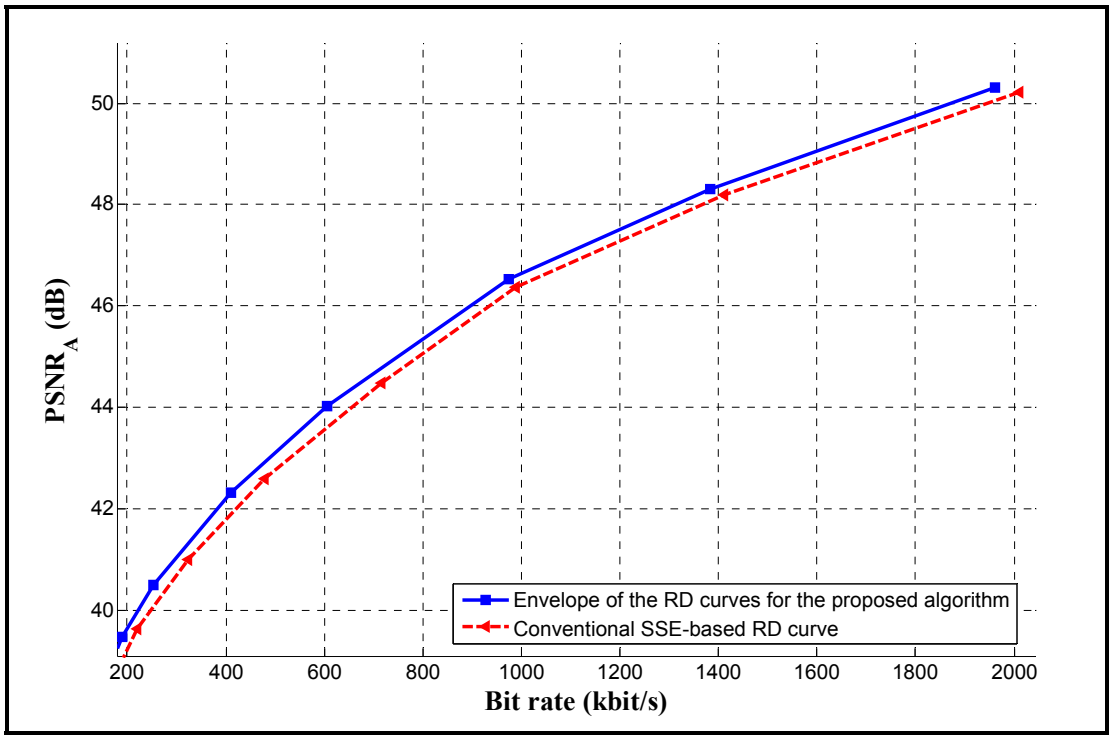

Figure 5.17 The RD curves for encoding 120 frames of "container" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

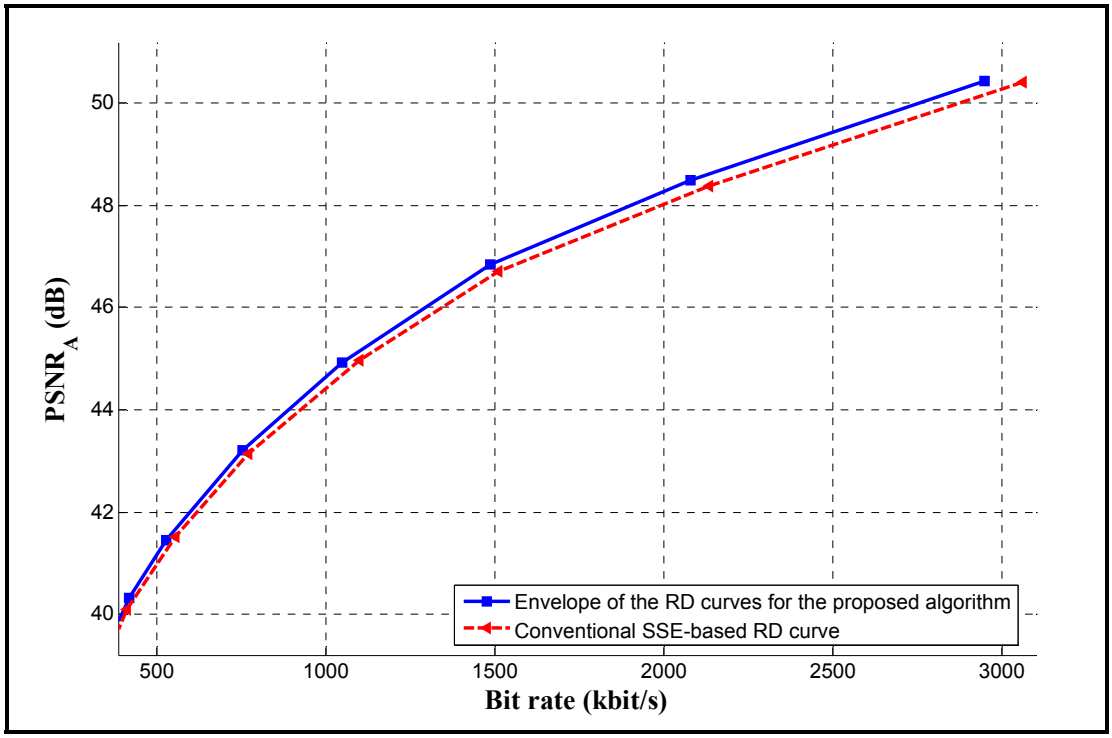

Figure 5.18 The RD curves for encoding 120 frames of "foreman" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

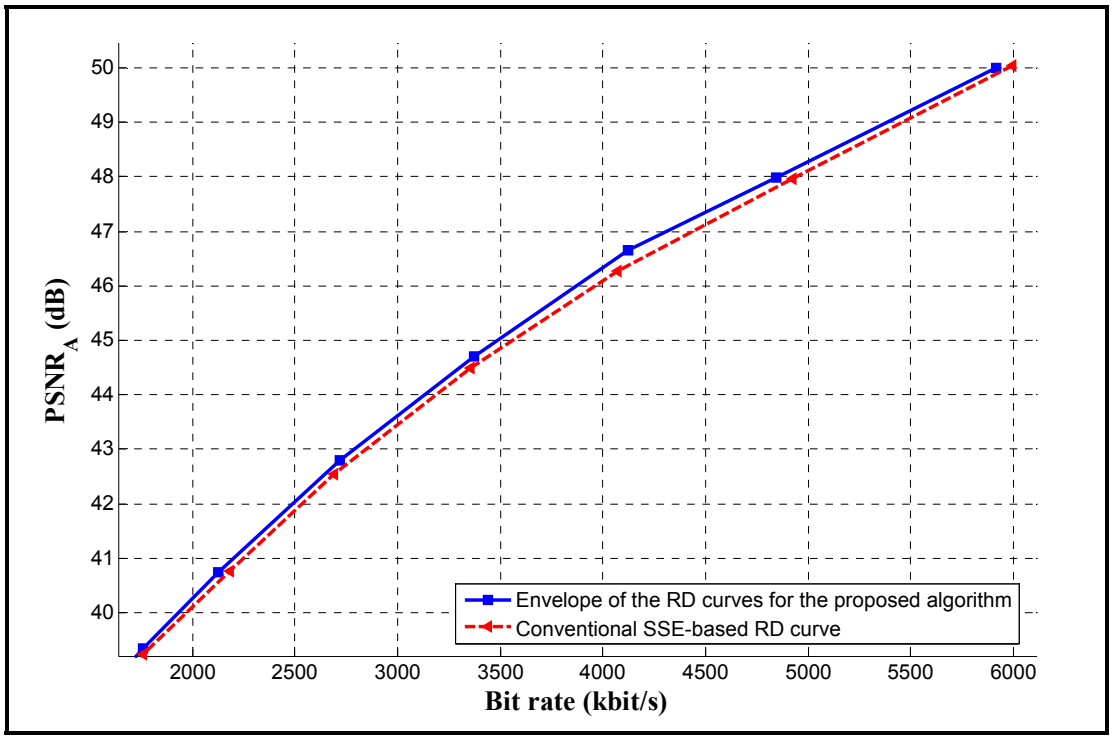

Figure 5.19 The RD curves for encoding 120 frames of "football" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

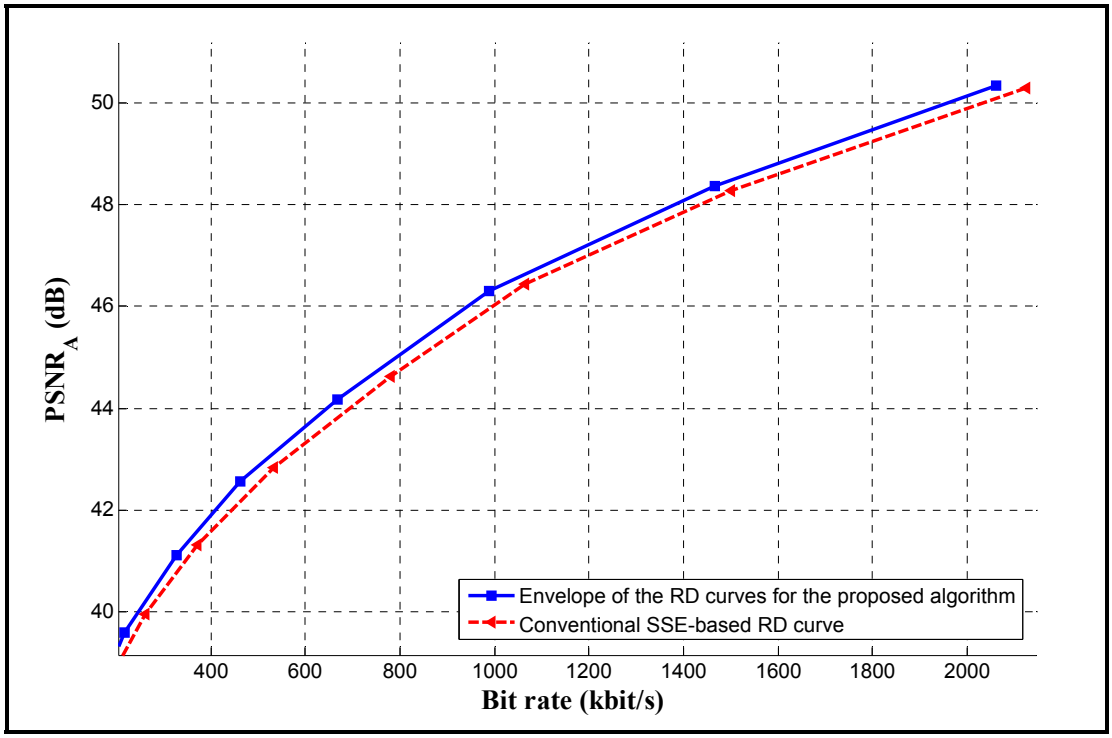

Figure 5.20 The RD curves for encoding 30 frames of "container" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

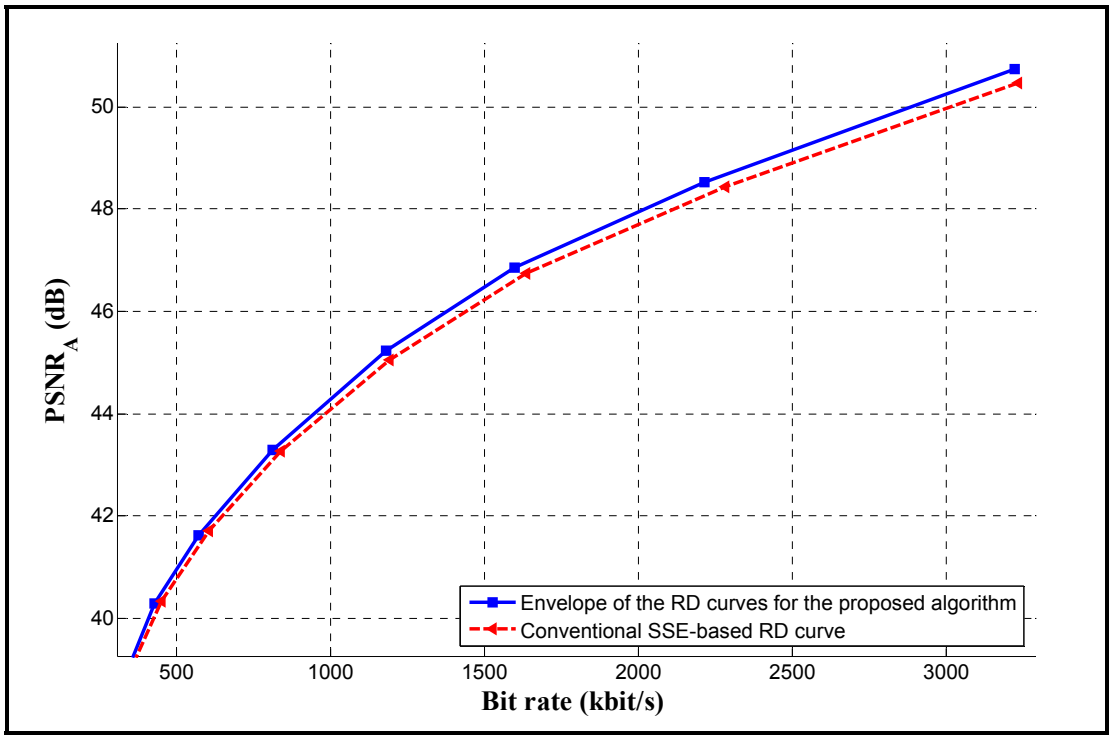

Figure 5.21 The RD curves for encoding 30 frames of "foreman" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

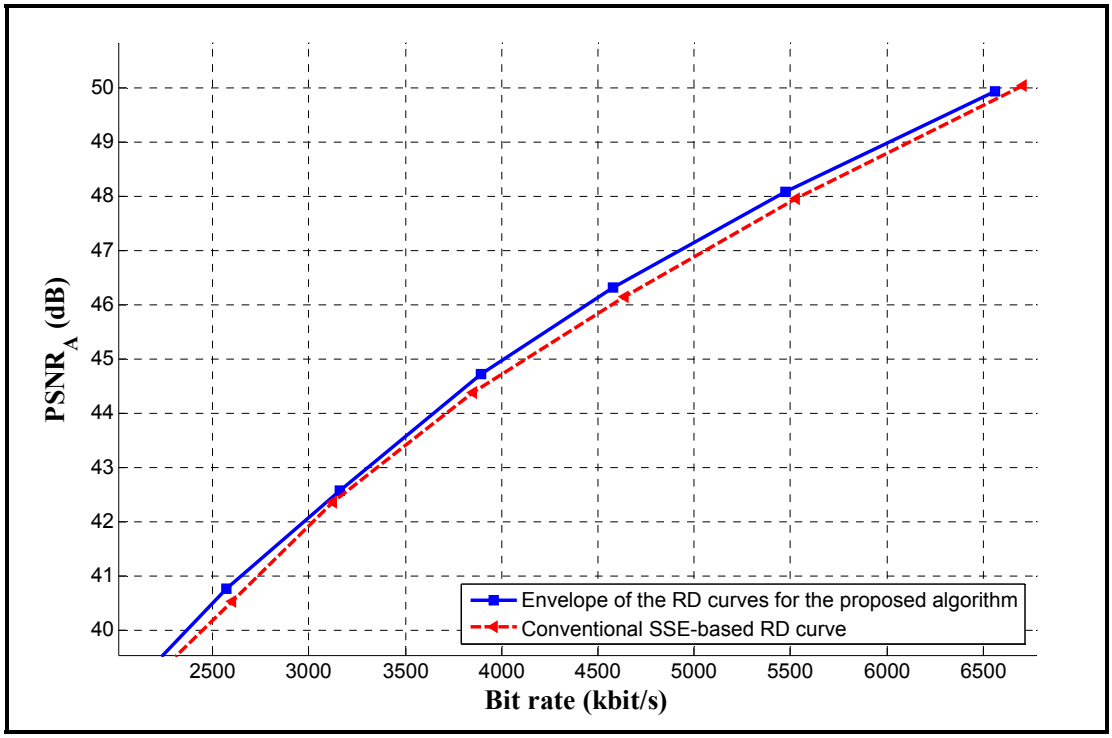

Figure 5.22 The RD curves for encoding 30 frames of "football" when $\lambda_{MOTION} \neq \lambda_{MODE}$.

5.4.3 RD curves for non-adapted Lagrange multiplier λ_p

The rate-distortion curves for SSE_A-based mode decision, shown in the previous subsections, generated with the adapted Lagrange multipliers using our search algorithms. In this section, we validate the accuracy of calculated Lagrange multipliers in table 5.7. To this end, we verify the efficiency of SSE_A-based mode decision by encoding the sequences using non-adapted Lagrange multipliers, i.e. the Lagrange multiplier values in table 5.7 are used for SSE_A-based mode decision. As observed in the previous subsections, the SSE_A-based mode decision is mainly efficient for slow sequences. Here, we tested the RD performance for two sequences: "akiyo" and "hall". The "akiyo" is a slow/simple video sequence but "hall" more complex relative to "akiyo". The RD curves for "akiyo" are shown in figure 5.23. It can be seen that for the sequence "akiyo" the SSE_A-based mode decision is still efficient when non-adapted Lagrange multipliers are used in the mode decision process, however, as expected, for the more complex sequence "hall" the performance is not that good (see figure 5.24).

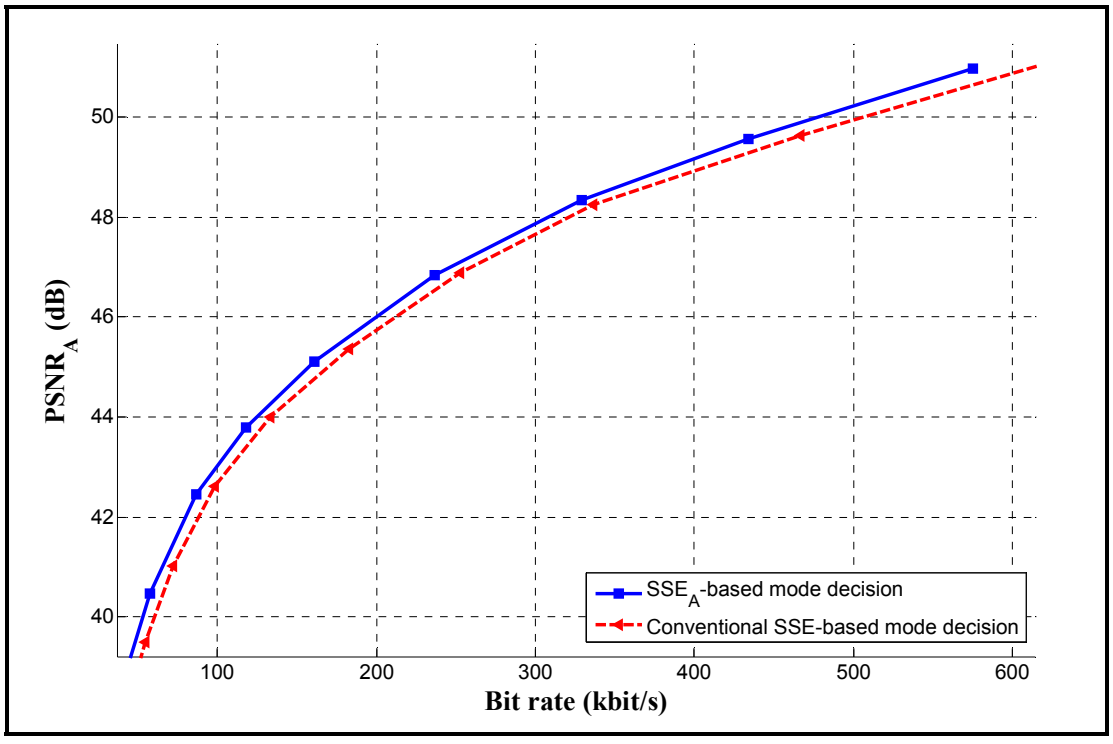

Figure 5.23 The RD curves for encoding 120 frames of "akiyo" when $\lambda_{MOTION} = \lambda_{MODE}$ and using non-adapted λ_p.

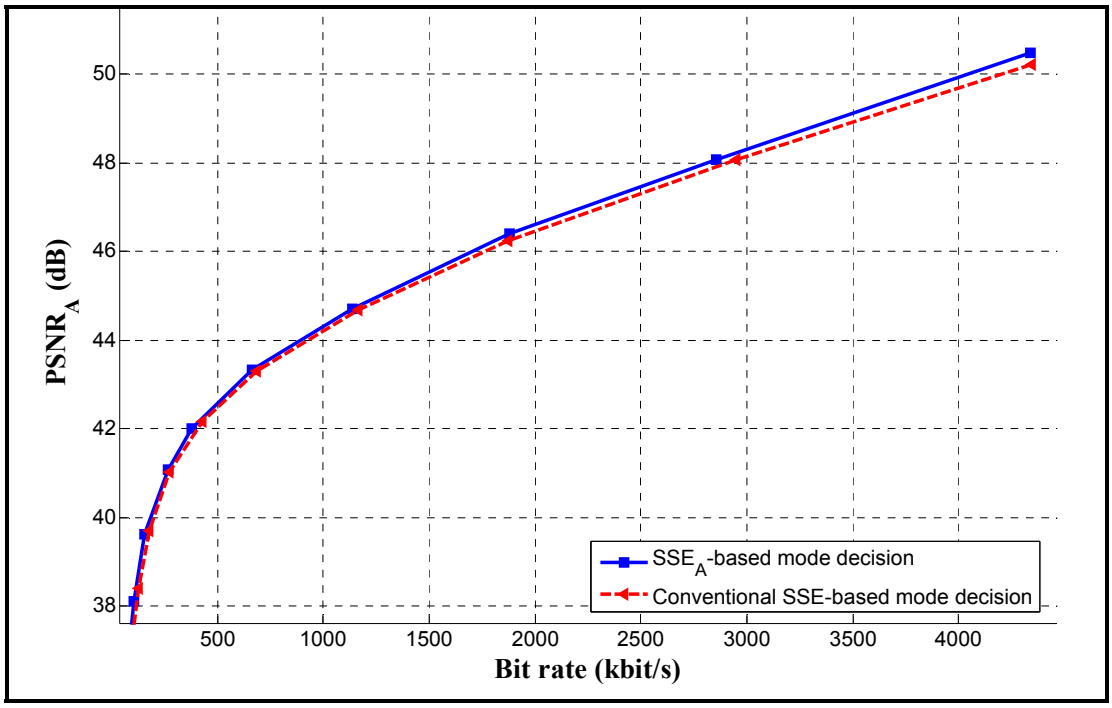

Figure 5.24 The RD curves for encoding 120 frames of "hall" when $\lambda_{MOTION} = \lambda_{MODE}$ and using non-adapted λ_p.

5.5 Conclusion and future research

In this thesis, we focused on mode decision in H.264/AVC as a potential usage of quality/distortion metrics. We applied the proposed metric SSE_A as a distortion metric in H.264/AVC RDO mode decision. By using SSE_A, the quality of each frame is optimized relative to $PSNR_A$. We first analyzed the relationship between macroblock distortions in pixel and wavelet domains theoretically and empirically. Our analysis showed that at low *QP*s (high rates) SSE_A have a very different behaviour from SSE.

We implemented SSE_A within the JM reference software, and proposed a search algorithm to determine the Lagrange multiplier for each individual *QP*. Applying the search algorithm on different video sequences showed that after a certain *QP* value, the Lagrange multiplier is changing exponentially with the change in *QP* according to the motion characteristics of video sequences. The RD curves confirmed that using the proposed SSE_A-based mode

decision method, instead of the traditional SSE-based mode selection, can reduce the bitrate about 5% at the same $PSNR_A$ for low to medium motion activity sequences.

For future research, we can repeat our experiments on sequences with larger resolutions, and use other visual quality metrics for building the RD curves. Finally, it would be helpful to express the Lagrange multiplier table (like table 5.7) as a mathematical formula.

CHAPTER 6

CONTRIBUTIONS

The main contributions of our research can be listed briefly as follows:

1. A novel wavelet-domain framework was proposed for FR quality assessment of images. By using our framework, we can calculate the visual quality not only with lower computational complexity, but also with higher prediction accuracy compared to the well-known methods. The proposed framework can be applied to both top-down and bottom-up approaches.

2. A formula was derived which gives the required number of wavelet decomposition levels in our framework based on viewing condition.

3. A novel contrast map function was proposed in the wavelet domain for pooling the quality/distortion maps of the metrics.

4. A very accurate structural similarity metric ($SSIM_{DWT}$) was proposed with a computational complexity lower than SSIM index, and tested on different image and video databases.

5. A method was proposed for low-complexity computation of visual information fidelity in the discrete wavelet domain. The proposed metric (VIF_{DWT}) is more accurate than the original VIF index, and its computational complexity is about 5% of it.

6. Using our framework, a PSNR-based metric created with a prediction accuracy much higher than the conventional PSNR, and a computational complexity comparable to the PSNR.

7. An error-based metric (AD_{DWT}) was proposed by applying our framework on absolute difference of images. This metric is competitive with top-down approaches in terms of prediction accuracy, yet its computational complexity is very low.

8. Prediction accuracy of different image quality metrics, including our proposed models, was evaluated on different image and video databases, and compared together using different statistical measures.

9. The developed metric SSE_A was used as distortion measure instead of SSE for RDO-based mode decision in video encoding. This metric was implemented in H.264 reference software. The relationship between SSE and SSE_A was discussed and we analyzed both theoretically and empirically.

10. A search algorithm was described to determine the mode decision Lagrange multiplier at each *QP* for the proposed distortion measure SSE_A. Using the proposed search algorithm, the Lagrange multiplier values tabulated for three classes of videos, i.e. high, medium, and low motion activity sequences.

It is worth mentioning that this work has been published in two journal papers (Rezazadeh and Coulombe, 2013d), (Rezazadeh and Coulombe, 2012b), three conference proceeding (Rezazadeh and Coulombe, 2009), (Rezazadeh and Coulombe, 2010), (Rezazadeh and Coulombe, 2011), and led to three granted patents (Rezazadeh and Coulombe, 2013a), (Rezazadeh and Coulombe, 2013b), (Rezazadeh and Coulombe, 2013c), (Rezazadeh and Coulombe, 2012a) and one patent application (Coulombe and Rezazadeh, 2012).

CONCLUSION

In this thesis, we presented a novel framework for full reference quality assessment of images (or video frames). In our quality assessment framework, the approximation subband of decomposed images has a key role in both prediction accuracy improvement and complexity reduction. Therefore, this framework works in the discrete wavelet domain, and specifically makes use of the Haar wavelet for its distinctive characteristics. We showed that the proposed framework can be applied to both categories of top-down approaches and the error-based methods. A formula was derived according to the principles of the HVS to calculate the appropriate number of wavelet decomposition levels for the error-based techniques. For the techniques that generate a quality (or distortion) map, a contrast map function was defined in the discrete wavelet domain for pooling the quality map. Another unique feature of our framework is the introduction of a low-complexity edge map for each image, however based on the potential application we can choose to use or not the edge map to reduce furthermore the computational complexity.

Using the proposed framework, we introduced four different quality assessment methods, including $SSIM_{DWT}$, VIF_{DWT}, $PSNR_{DWT}$, and AD_{DWT}. These metrics offer different levels of computational complexity and prediction accuracy. The performance of the proposed methods evaluated on three different image and video test databases: the LIVE image database, the TID2008 image database, and the LIVE video database. Different statistical measures such as the linear correlation coefficient, Spearman and Kendall rank correlation coefficients, and RMSE were employed to evaluate the performance of objective quality models. Furthermore, the computational complexity of various quality metrics was analyzed and compared to the H.264 encoding complexity. Performance evaluation of metrics on the LIVE image database showed that VIF_{DWT} is more accurate than the original VIF index, while its computational complexity is about 5% of it. It is notable that the performance of quality metrics may slightly change from one test database to another database. For example, the proposed AD_{DWT} shows the best performance over the examined distortion types of the TID2008 image database, however the performances of the other proposed quality metrics,

including $SSIM_{DWT}$ and VIF_{DWT}, are very close to it. Simulations on the LIVE video database illustrated that the $SSIM_{DWT}$ is the most accurate metric for visual quality assessment of H.264 compressed videos. Moreover, using the framework we proposed a more accurate version of the conventional PSNR, i.e. $PSNR_{DWT}$, which can be computed with a complexity nearly the same as PSNR (when the edge map is not considered).

An appealing feature of the framework is that it does not introduce many extra parameters in addition to those belonging to the original versions of the metrics. This feature made it possible to use the metrics with the same parameter settings across different databases. Since the proposed quality assessment framework provides a very good trade-off between accuracy and complexity, it can be used in many image/video processing applications. For future studies, it is worth developing techniques for temporal pooling of our proposed quality metrics for low-complexity perceptual video quality assessment.

In the second part of the thesis, we investigated the role of quality metrics in the video encoding mode decision. We first studied different approaches of mode decision for the H.264 coding standard. Then, our proposed metric, i.e. SSE_A, was used as the distortion measure inside the H.264/AVC mode decision process. To determine the Lagrange multiplier for the implemented metric, an exhaustive search algorithm was proposed. The search algorithm was applied on three different types of test sequences and its results tabulated. It was observed that the Lagrange multiplier values as a function of *QP* vary exponentially based on the motion features of the videos. We evaluated the performance of our SSE_A-based mode decision algorithm by simulating the rate-distortion (RD) curves of encoding process. The RD curves showed that at the same $PSNR_A$, the SSE_A-based mode decision offers 5% bitrate reduction, on average, relative to the conventional SSE-based method for low motion activity and medium motion activity sequences. It should be noted that the proposed mode decision algorithm has the same computational complexity as the conventional SSE-based method and hence, it does not impose any additional computational complexity to the encoding process. Finally, we remark that for future research it would be valuable to

optimize the coding mode decision process in terms of other quality metrics, and build the RD curves using them.

ANNEX I

BOTTOM-UP APPROACHES FOR VISUAL QUALITY ASSESSMENT

The visible difference predictor (VDP) model proposed by Daly in (Daly, 1992) is one of the most general and elaborate image quality metrics in the literature. In this model, a variation of Watson's cortex transform, as shown in figure-A I-1, is used to decompose the image into five spatial levels followed by six orientation levels. Then, a threshold map is computed for each channel from the contrast in that channel. The Daly model accounts for a number of processing stages, including a point-wise nonlinearity, spatial frequency filtering by contrast sensitivity function (CSF), a channel decomposition, contrast calculation, masking calculation, and a probability of detection calculation. To account for contrast sensitivity, the VDP filters the image by the CSF before the frequency decomposition. Once this normalization is accomplished to account for the varying sensitivities of the HVS to different spatial frequencies, the thresholds derived in the contrast masking stage become the same for all frequencies. A distinct feature of Daly model is that not only the reference image but also the distorted image are included in the calculation of masking factor. In the final error pooling stage, a psychometric function is used to compute the probability of discrimination at each pixel of the reference and test images to obtain a spatial map.

Lubin's model, which is also known as the Sarnoff visual discrimination model (VDM) (Lubin, 1995), is another model that attempts to estimate the probability of detection of the differences between the reference image and distorted image. Preprocessing steps in this model include calibration for distance of the observer from the images. In this model, the images are decomposed using a Laplacian pyramid into seven resolutions (radial frequency bands) after low-pass filtering and resampling. To reflect orientation selectivity, each pyramid level is then decomposed into four orientations using a bank of steerable filters. The frequency decomposition so obtained is illustrated in figure-A I-2. Further details of this algorithm can be found in (Lubin, 1995).

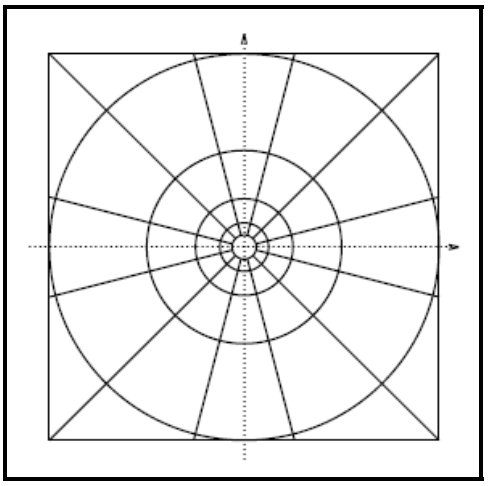

Figure-A I-1 The decomposition of the frequency plane corresponding to Cortex transform (Daly). The range of each axis is from $-u_s/2$ to $+ u_s/2$ cycles per degree (u_s is the sampling frequency).
(Adapted from (Bovik, 2009))

In the Teo and Heeger model (Teo and Heeger, 1994), the channel decomposition is applied after a front-end linear filtering stage. This model adopts a steerable pyramid decomposition with six orientations, which is a polar separable wavelet design that avoids aliasing in the subbands. As opposed to the other models, the normalization process in this model is calculated for the reference and distorted images separately before a squared error signal is computed. It's worth noting that the quality assessment models discussed so far use transforms which are polar separable, and belong to a category of decompositions that are mimicking processing in the visual cortex. In following, we introduce some other quality assessment models utilizing transforms that are often used in compression systems (image coders) and model the thresholds of visibility for each of the channels.

The Safranek-Johnston model was one of the first image coders designed for perceptual image coding (Safranek and Johnston, 1989). It is calibrated for a given CRT display and viewing conditions (six times image height). It decomposes the image signals using a separable generalized quadrature mirror filter (GQMF) bank for subband analysis/synthesis. This transform is invertible, such that it can be used for both analysis and synthesis. The perceptual model specifies the amount of noise that can be added to each subband of a given

image so that the difference between the output image and the original is just noticeable. A brightness adjustment identical for all subbands is also included. The overall normalization factor for a coefficient is computed as the product of the baseline sensitivity factor, the brightness adjustment factor, and the masking factor. At the final stage, the Minkowski metric is used for error pooling.

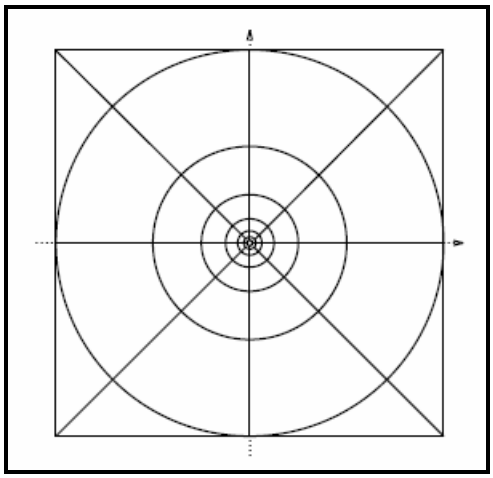

Figure-A I-2 The decomposition of the frequency plane corresponding to Lubin's transform. The range of each axis is from $-u_s/2$ to $+ u_s/2$ cycles per degree (u_s is the sampling frequency).
(Adapted from (Bovik, 2009))

Many compression standards are based on a discrete cosine transform (DCT) decomposition. The DCT is a variation of the discrete Fourier transform that partitions the frequency spectrum into uniform subbands. Watson (Watson, 1993) presented a model known as DCTune that computes the visibility thresholds for the DCT coefficients, and thus provides a metric for image quality. Watson's model was developed as a means to compute the perceptually optimal image-dependent quantization matrix for DCT-based image coders like JPEG. Watson's DCT model first divides the original reference and degraded images into distinct 8×8 blocks, and a visibility threshold is calculated for each coefficient in each block. In this model, three factors determine the visibility threshold: the baseline contrast sensitivity threshold, luminance masking, and contrast/texture masking. The errors between the reference image and distorted image are normalized using the visibility threshold.

The VSNR (Chandler and Hemami, 2007) is another advanced general purpose HVS-based metric that quantifies the visual fidelity of natural images based on near-threshold and suprathreshold properties of human vision. Psychophysical experiments used in the VSNR to quantify the visual detectability of distortions in natural images. As opposed to most other models discussed, the VSNR attempts to capture a mid-level property of the HVS known as global precedence. Viewing conditions are taken into account in the preprocessing stage by modeling the display characteristics, and considering the viewing distance and the spatial resolution of display. After preprocessing, both reference image and the errors between the reference and distorted images are decomposed into five levels using discrete wavelet transform (DWT). The 9/7 biorthogonal filters are used in DWT decomposition. Then, it computes the contrast detection threshold to assess the detectability of the distortions for each subband of the wavelet decomposition.

It's worth mentioning that some other well-known methods of this category exploit Fourier transform rather than multiresolution decomposition, such as wSNR, NQM (Damera-Venkata *et al.*, 2000) and PQS (Miyahara, Kotani and Algazi, 1998).

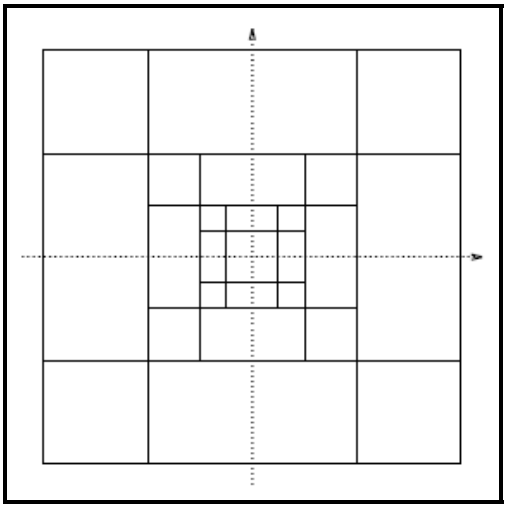

Figure-A I-3 The decomposition of the frequency plane corresponding to wavelet transform. The range of each axis is from $-u_s/2$ to $+ u_s/2$ cycles per degree (u_s is the sampling frequency).
(Adapted from (Bovik, 2009))

ANNEX II

FAST HIGH COMPLEXITY MODE RDO AND LOW COMPLEXITY MODE DECISION

In fact, the goal of these low complexity and fast high complexity methods is to achieve good rate-distortion performance at a reduced computational cost. Some of the fast coding mode selection methods have been studied and suggested fast motion estimation algorithms to reduce the computation cost, like (Tourapis, Cheong and Topiwala, 2005), which have been adopted in the JM reference software too (Joint Video Team (JVT) H.264/AVC Reference Software). Other studies have investigated fast fractional pixel search (Chen *et al.*, 2002), and fast intra prediction mode decision, such as (Pan *et al.*, 2003), (Tsukuba *et al.*, 2005), and (Quan and Ho, 2010).

Under the fast high-complexity mode, different approaches are exploited for fast inter mode decision, to choose the best macroblock mode in a computationally efficient way. One of the popular methods is early SKIP mode decision in inter-coded slices to check if the given macroblock is likely to choose the SKIP mode as the best mode. Since the number of skipped macroblocks increases with the increase of the quantization parameter, a mode decision algorithm can wisely select macroblocks as the SKIP mode before checking all modes. So, a large portion of computation saving is obtained. The fast mode decision method in (Jeon and Lee, 2003) checks the early SKIP mode decision after applying motion search with 16×16 block size, and the boundary error is used to decide the inter or intra mode for a macroblock. Reliable features of current macroblock with light computation cost or reusing features of correlated adjacent macroblocks are desirable in fast mode decision design. The general computation procedure for fast high-complexity mode decision is shown in figure-A II-1 (Lim, Sullivan and Wiegand, 2005).

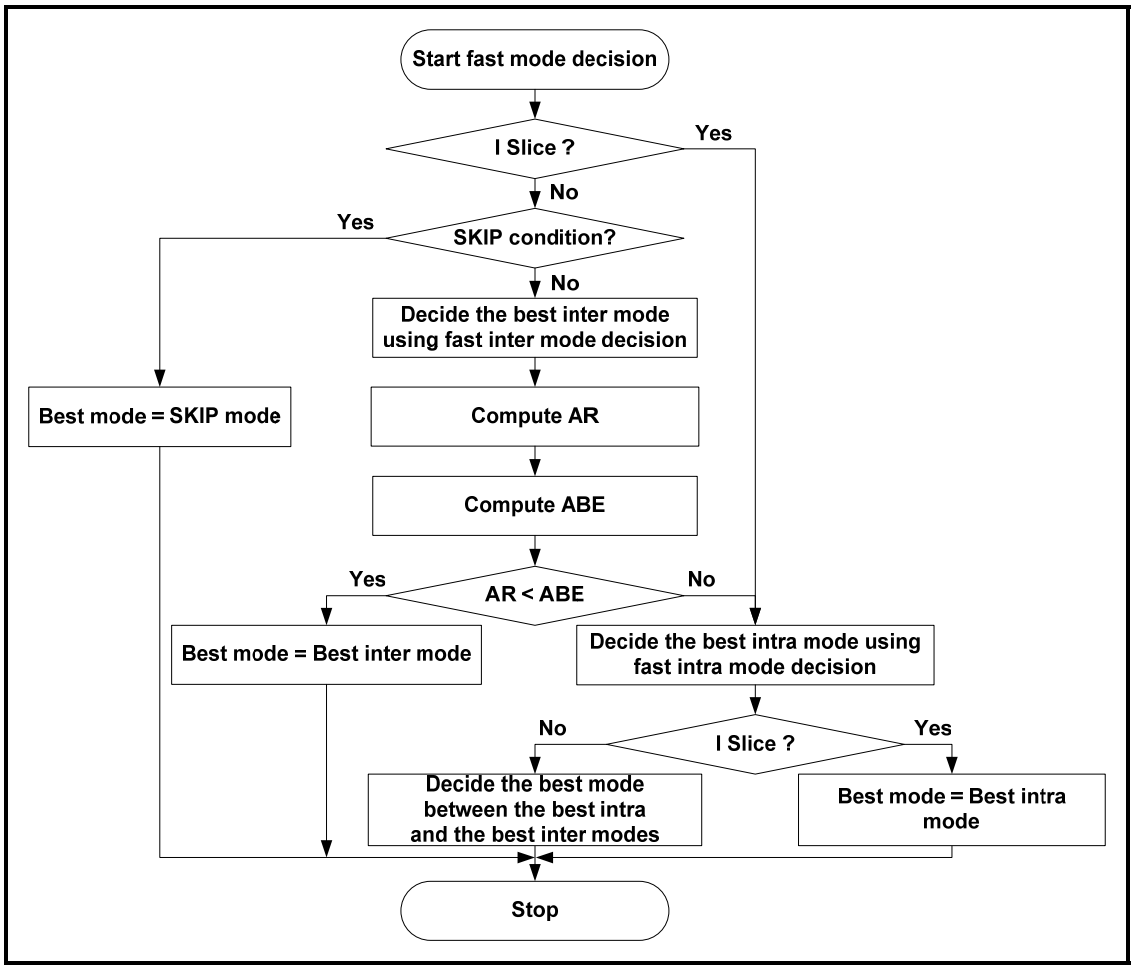

Figure-A II-1 Flow chart of macroblock mode decision under fast high-complexity mode. (Adapted from (Lim, Sullivan and Wiegand, 2005))

where ABE denotes the average error between pixels at the boundary of the current and its adjacent encoded blocks, and AR is the average number of bits consumed to encode the motion compensated residual data under the best inter mode (Lim, Sullivan and Wiegand, 2005).

Authors in (Choi, Lee and Jeon, 2006) have proposed two methods similar to approach in (Jeon and Lee, 2003) to reduce the computational complexity of mode decision process: early SKIP mode decision and selective intra mode decision. Simulation results show that their proposed methods can significantly reduce the entire encoding time by about 60% with only negligible coding loss. In (Lin, Fink and Bellers, 2007), an H.264 fast mode decision

algorithm is proposed that reduces computations by using a statistical dependency of macroblock RD costs. A number of motion estimation and/or intra prediction mode decision parts are skipped in this algorithm with the help of an adaptive threshold. It is demonstrated that macroblock RD costs of a given mode have high correlation to those of the temporal and spatial adjacent macroblocks, especially in the previous co-located macroblock. By adopting statistical dependency, an almost 2x speedup is gained in the proposed fast mode decision algorithm. In (Bystrom, Richardson and Zhao, 2008) a method is introduced for rapid SKIP mode decision by using a Bayesian framework. The rate-distortion cost difference between coding and skipping a macroblock is used as the single decision feature and an appropriate decision threshold determined following modeling of the cost difference's class-conditional PDFs. The threshold's parameters are modeled as functions of the quantization parameter and a local sequence activity factor as measured by frame difference. In view of the fact that a large percentage of macroblocks are skipped in low-motion sequences, the proposed method is particularly efficient for low-activity sequences and/or at lower bitrates. Simulation results show that the proposed approach can achieve a time savings of over 80% for low-motion sequences at a negligible decrease or, sometimes, a slight increase in quality compared to reference H.264 codec.

To avoid the expensive computation of Lagrange costs, transform-domain bit-rate estimation and distortion measures are proposed in (Tu, Yang and Sun, 2006) based on quantized and inverse quantized integer transform coefficients, for the inter-mode decision in H.264/AVC encoding. The bit-rate is efficiently estimated by modeling coded bits consumption as a function of the number and the levels of the nonzero quantized transform coefficients. By the proposed scheme in (Tu, Yang and Sun, 2006), entropy coding, inverse transform, and pixel reconstructions (block reconstruction) are not required in the process and can be skipped. With simulations, it is demonstrated that the proposed RD estimation method can achieve about 40% reduced time of computing the RD cost for the inter mode decision, and saves about 17% total encoding time with ignorable degradation in coding performance comparing with original RD optimized H.264/AVC encoder.

In (Wang, Kwong and Kok, 2007) an efficient algorithm is presented to jointly optimize mode decision and motion estimation. The method use theoretical analysis to study the sufficient condition to detect all-zero blocks in H.264, and consequently adaptive thresholds are obtained to early terminate mode decision and motion estimation. Additionally, the proposed algorithm introduces temporal-spatial checking (TSC), thresholds based prediction (TBP), and monotonic error surface based prediction (MESBP) methods to further skip checking unnecessary modes. Authors in (Shen *et al.*, 2008) propose a fast inter mode decision to choose the best prediction mode utilizing the spatial continuity of motion field, which is generated by performing motion estimation on 4×4 block size using the nearest reference frame. The method assumes that the continuously moving macroblocks in the frame do not require to be further split into smaller blocks. Therefore, by predicting where motion is continuous, significant computational savings can be achieved for the motion estimation and RD cost computations for small sizes. The Sobel operator is applied to both, horizontal and vertical, components of the motion field to detect motion continuity. It is shown by simulations that the algorithm can save more than 50% computational complexity, with negligible loss of coding efficiency compared to the full mode decision.

There are many more various techniques for fast high complexity mode decision, however, in the next sections we focus on the high complexity mode and do our simulations based on that. That means, in the H.264 JM reference software, we set the parameter "RDOptimization" to 1.

ANNEX III

STATISTICS OF MACROBLOCK DISTORTIONS BETWEEN THE PIXEL DOMAIN AND WAVELET DOMAIN FOR THE SEQUENCE "MOBILE"

In this annex, we bring the complementary results for the sequence "mobile", including the tables and scatter plots, which show the relationship between the pixel domain and wavelet domain macroblock distortion measures through different statistical metrics.

Table-A III-1 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the sequence "mobile" (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock, and then metrics' values averaged for each frame over the whole sequence.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.9820	0.9317	0.9945
	SRCC	0.8090	0.7917	0.9943
	KRCC	0.6124	0.6145	0.9467
	α	0.2904	0.3981	0.7243
	β	-6.3287	-578.8944	-2.0996e+004
MSE vs MSE_A	LCC	0.9820	0.9317	0.9945
	SRCC	0.8090	0.7917	0.9943
	KRCC	0.6124	0.6145	0.9467
	α	1.1615	1.5925	2.8972
	β	-0.0989	-9.0452	-328.0611
PSNR vs $PSNR_A$	LCC	0.9850	0.9271	0.9906
	SRCC	0.8448	0.7921	0.9910
	KRCC	0.6509	0.6299	0.9325
	α	1.0708	1.1584	1.3523
	β	2.3610	-0.6631	-5.3376

Table-A III-2 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the frame number 25 of the sequence “mobile” (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.7058	0.8411	0.9068
	SRCC	0.6744	0.8016	0.9007
	KRCC	0.4878	0.6126	0.7360
	α	0.1786	0.2469	0.4573
	β	58.5080	937.6635	4.7954e+003
MSE vs MSE_A	LCC	0.7058	0.8411	0.9068
	SRCC	0.6744	0.8016	0.9007
	KRCC	0.4878	0.6126	0.7360
	α	0.7145	0.9875	1.8290
	β	0.9142	14.6510	74.9275
PSNR vs $PSNR_A$	LCC	0.7291	0.9255	0.9659
	SRCC	0.6744	0.8016	0.9007
	KRCC	0.4880	0.6126	0.7360
	α	0.5853	0.7310	0.9119
	β	24.0301	13.3117	4.9387

Table-A III-3 The relationship between pixel domain and wavelet domain metrics through different statistical methods for the frame number 75 of the sequence “mobile” (CIF, 30Hz). The distortion/quality metrics calculated for each 16×16 macroblock.

		***QP* = 16**	***QP* = 30**	***QP* = 44**
SSE vs SSE_A	LCC	0.6483	0.8433	0.9118
	SRCC	0.6256	0.8098	0.9087
	KRCC	0.4494	0.6179	0.7450
	α	0.1935	0.2385	0.4822
	β	51.6224	968.6478	6.1837e+003
MSE vs MSE_A	LCC	0.6483	0.8433	0.9118
	SRCC	0.6256	0.8098	0.9087
	KRCC	0.4494	0.6179	0.7450
	α	0.7739	0.9539	1.9287
	β	0.8066	15.1351	96.6203
PSNR vs $PSNR_A$	LCC	0.6837	0.9190	0.9691
	SRCC	0.6256	0.8098	0.9087
	KRCC	0.4494	0.6179	0.7450
	α	0.6482	0.7357	0.9087
	β	21.1963	13.2270	4.6907

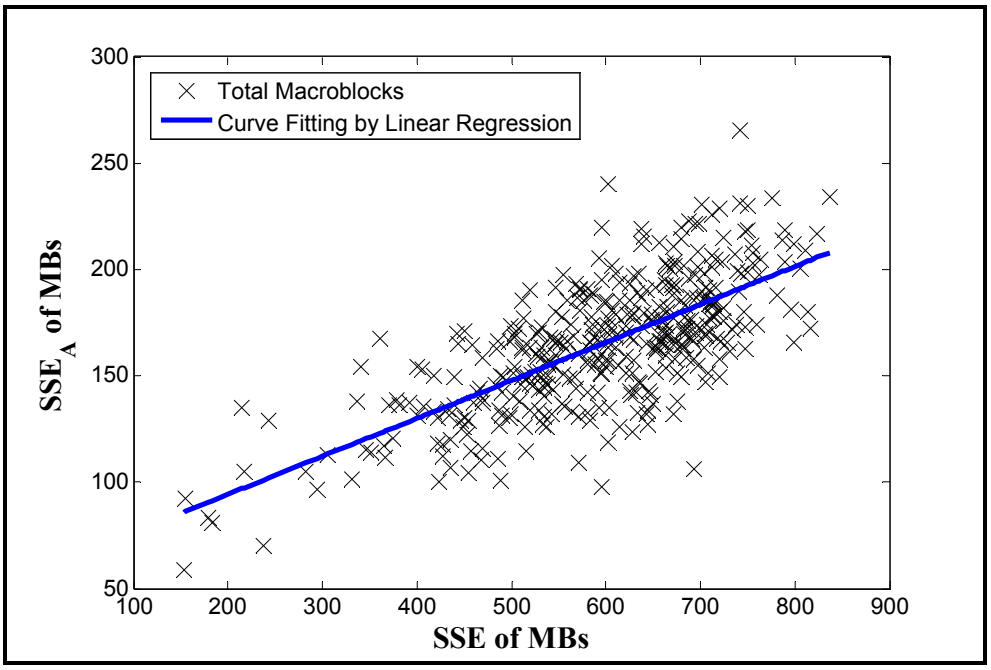

Figure-A III-1 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 16. The distortion metrics calculated for each 16×16 macroblock.

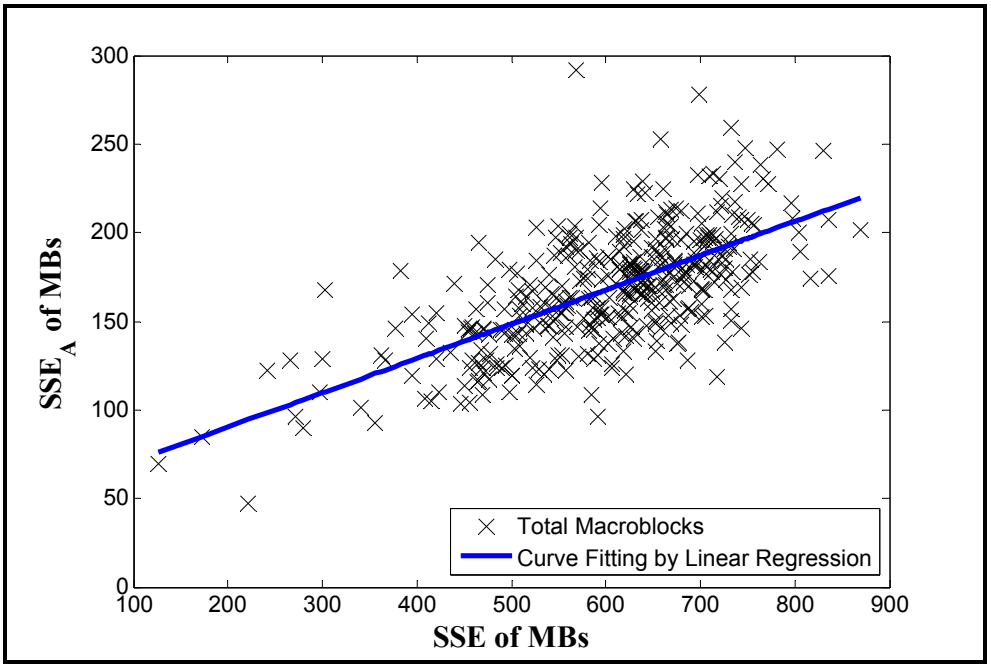

Figure-A III-2 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 16. The distortion metrics calculated for each 16×16 macroblock.

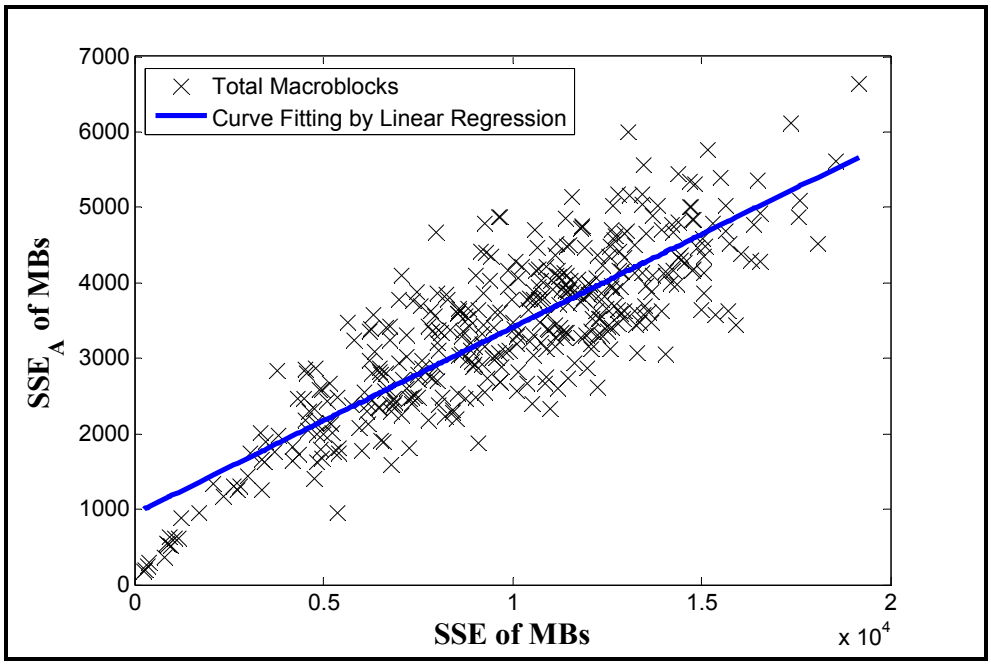

Figure-A III-3 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 30. The distortion metrics calculated for each 16×16 macroblock.

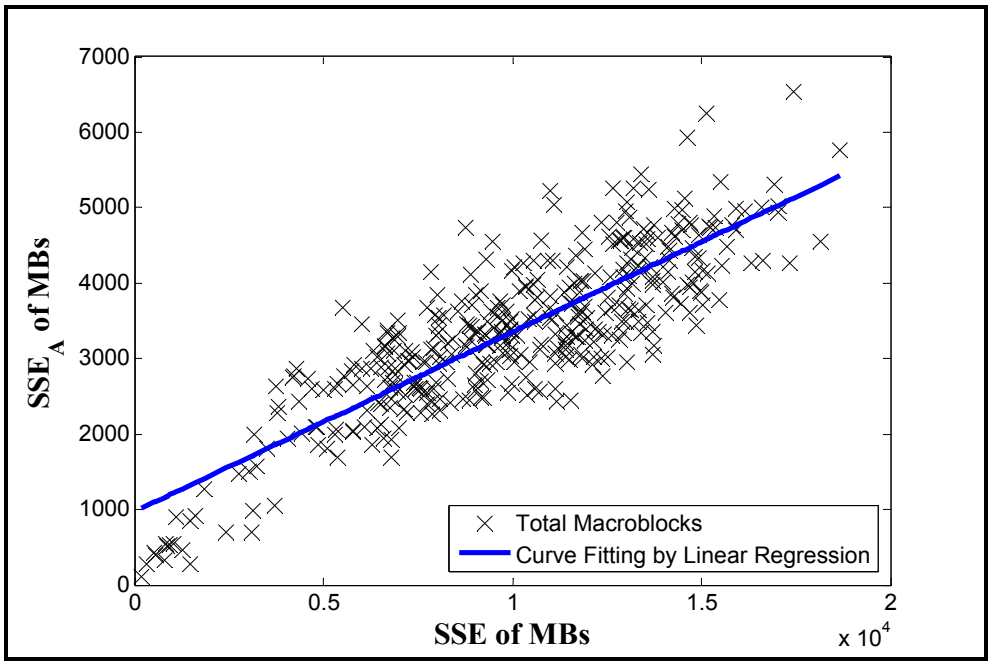

Figure-A III-4 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 30. The distortion metrics calculated for each 16×16 macroblock.

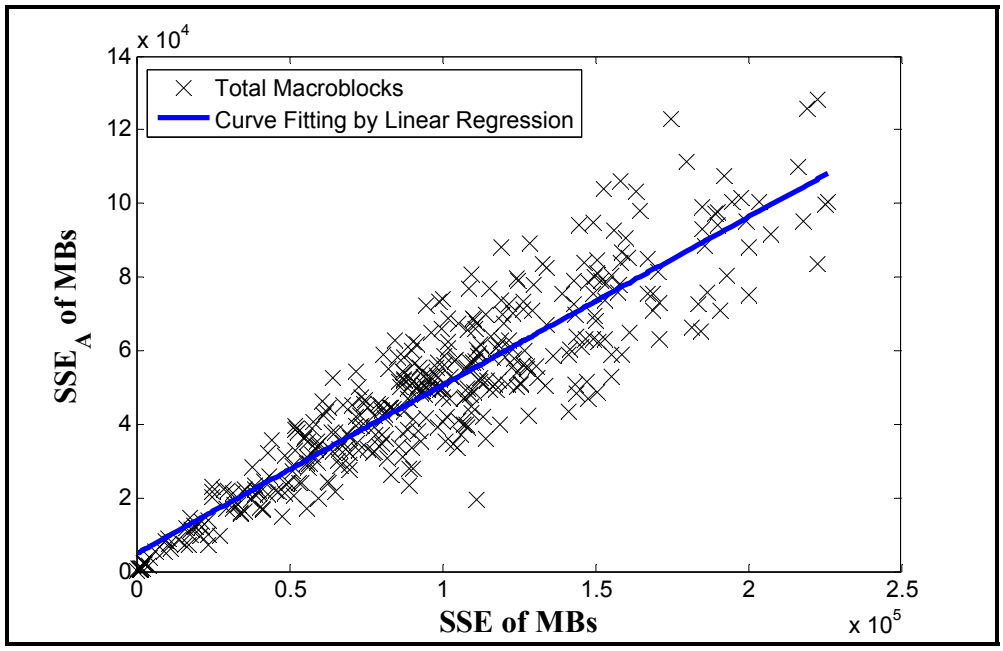

Figure-A III-5 Scatter plot of SSE vs. SSE_A for the frame number 25 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 44. The distortion metrics calculated for each 16×16 macroblock.

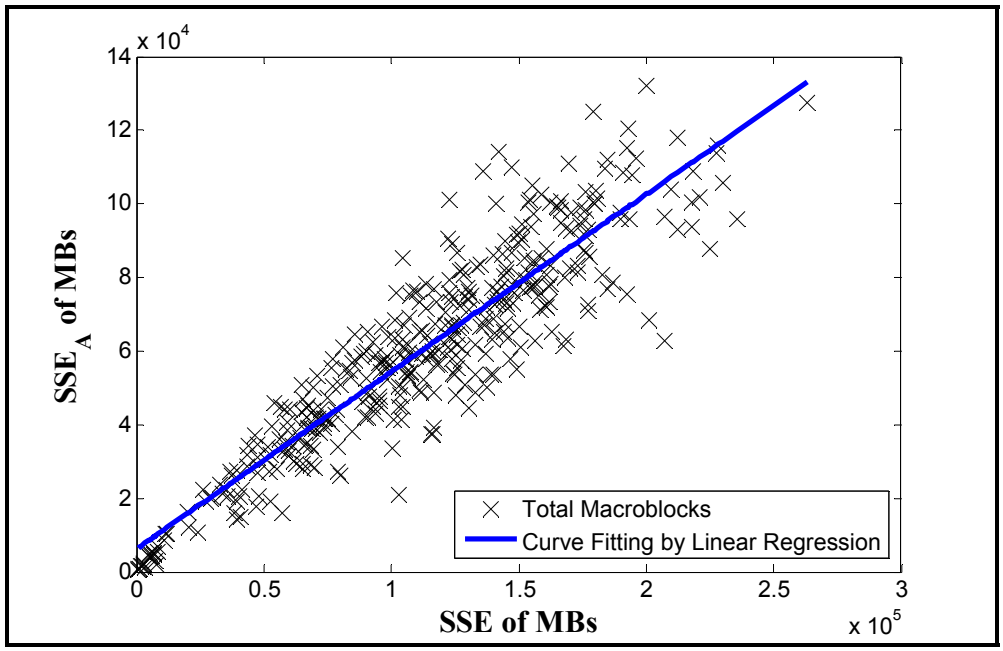

Figure-A III-6 Scatter plot of SSE vs. SSE_A for the frame number 75 of the sequence "mobile" (CIF, 30Hz) coded with *QP* = 44. The distortion metrics calculated for each 16×16 macroblock

ANNEX IV

MATLAB CODE FOR FITTING THE ENVELOPE TO A SET OF RD CURVES AND FINDING THE BEST POINT ON EACH OF THEM

```
clc
clear
close all


%%
%%%%%%%%%%%%%%%%%%% Load Original & Compressed RD Data %%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

S1 = load('/Original/container/container.mat');
R1_org = load('/Original/container/RateNumbersP.txt');
R1 = R1_org;

S2 = load('/Double/container/container.mat');
R2_org = load('/Double/container/RateNumbersP.txt');
R2 = R2_org;

%%
%%%%% Forming rate and Quality Vectors + Plot Wavelet Modified  Curves %%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

qpTot = 15;
jump = 58;

cnt=0;

for qp=16:2:44

    R2_T(cnt+1,:) = R2(1+(cnt*jump) : jump*(cnt+1));
    S2_PSNR_T(cnt+1,:)   = S2.QualityScoreFinal_Y_PSNR( 1+(cnt*jump) : jump*(cnt+1));
    S2_PSNR_A_T(cnt+1,:) = S2.QualityScoreFinal_Y_PSNR_A( 1+(cnt*jump) : jump*(cnt+1));
    S2_SSIM_T(cnt+1,:)   = S2.QualityScoreFinal_Y_SSIM( 1+(cnt*jump) : jump*(cnt+1));

    plot(R2_T(cnt+1,:), S2_PSNR_A_T(cnt+1,:), '-
ko','LineWidth',2.5,'MarkerEdgeColor','k','MarkerFaceColor','k','MarkerSize',5)

    hold on

    cnt = cnt+1;
end


%%
%%%%%%%%%%%%%%%%%%%%%%%% Interpolation Method %%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

cntSect = 0;

for qp=16:2:44

    cntSect = cntSect+1;

    if (cntSect == 1)

        Validity_Mat = ones(2,jump);
        for qpN = cntSect:(cntSect+1)
```

```
            for lam=1:jump
                R_Comp_T = R2_T(qpN,lam);
                Q_Comp_T = S2_PSNR_A_T(qpN,lam);

                for ii = cntSect:(cntSect+1)
                    for jj=1:jump

                        if (((R2_T(ii,jj) >= R_Comp_T) && (S2_PSNR_A_T(ii,jj) < Q_Comp_T)) ||
((R2_T(ii,jj) > R_Comp_T) && (S2_PSNR_A_T(ii,jj) <= Q_Comp_T)))
                            Validity_Mat(ii,jj) = 0;
                        end
                    end
                end

            end

        end


        [Ds,Dp] = p2gdist(R2_T(cntSect,:), S2_PSNR_A_T(cntSect,:), R1,
S1.QualityScoreFinal_Y_PSNR_A);

        [B,IX] = sort(Ds);

        M(cntSect) = 0;
        for i=1:length(IX)
            if ( Validity_Mat(cntSect,IX(i)) == 1 )
                M(cntSect) = IX(i);
                break
            end
        end

        if (M(cntSect) == 0)
            error('There is no valid point (lambda) with lower rate for this QP')
        end


        Rint(cntSect) = R2_T(cntSect, M(cntSect));
        Qual(cntSect) = S2_PSNR_A_T(cntSect, M(cntSect));

    %%
    elseif ((cntSect > 1)&&(M(cntSect-1) ~= 0))

        X1 = Rint(cntSect-1);
        Y1 = Qual(cntSect-1);

        for i=1:jump
            X2 = R2_T(cntSect, i);
            Y2 = S2_PSNR_A_T(cntSect, i);
            Ang(i) = (Y2-Y1)/(X2-X1);
        end


        %%%%%%%%%%%%%%%% Sorting of Angle to Choose the Minimum %%%%%%%%%%

        [B2,IX2] = sort(Ang);
        M(cntSect) = 0;
        for i=1:length(IX2)
            if ( ( R2_T(cntSect,IX2(i)) < Rint(cntSect-1)) && ( S2_PSNR_A_T(cntSect,IX2(i)) <
Qual(cntSect-1)) )
                M(cntSect) = IX2(i);
                break
            end
        end


        if (M(cntSect) == 0)
```

```
            str = sprintf('There is no valid point (lambda) with lower rate for the QP
#%d',cntSect);
            disp(str);
        end

        %%%%%%%%%%%%%%%%%%%%%  Angle with Threshold  %%%%%%%%%%%%%%%%%%%%%

        if (M(cntSect) ~= 0)

            AngThr = 0.0001;   % Threshold for the angle
            ArrayAng = find( (Ang <= (Ang(M(cntSect))+AngThr)) & (Ang >= (Ang(M(cntSect))-
AngThr)));

            for i=1:length(ArrayAng)

                if ( ( R2_T(cntSect,ArrayAng(i)) < Rint(cntSect-1)) && (
S2_PSNR_A_T(cntSect,ArrayAng(i)) < Qual(cntSect-1)) && (R2_T(cntSect,ArrayAng(i)) <
R2_T(cntSect,M(cntSect)) ) )

                    M(cntSect) = ArrayAng(i);

                end

            end


        end



        if (M(cntSect) ~= 0)
            Rint(cntSect) = R2_T(cntSect, M(cntSect));
            Qual(cntSect) = S2_PSNR_A_T(cntSect, M(cntSect));
        end




    else
        break;




    end % if (cntSect == 1)

end   % qp=16:2:44


plot(Rint,Qual,'-
gs','LineWidth',2.5,'MarkerEdgeColor','r','MarkerFaceColor','r','MarkerSize',5)
hold on

%%
%%%%%%%%%%%%%%%%%%%%% Plot Original R-D Curves %%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


plot(R1,S1.QualityScoreFinal_Y_PSNR_A,'--
bs','LineWidth',2.5,'MarkerEdgeColor','b','MarkerFaceColor','b','MarkerSize',5)

xlabel('Bit rate (kbit/s)  @ 30 Hz')
ylabel('PSNR_A (dB)')
grid on
%%
%%%%%%%%%%%%%%%%%%%%  Extracting Lambda Values   %%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
QP_vec = 16:2:44;   % 15 different values
lambda_start = 0.05;
lambda_end = 20000;


Lambda = [];
i = 1;
Lambda(i) = lambda_start;

while ( 1.25*Lambda(i) <= lambda_end )

    i = i+1;
    Lambda(i) = Lambda(i-1) * 1.25;
end

disp('Selected Lambda Numbers Are = ')
disp(M);

disp('Number of USED QP = ')
disp(length(M));

disp('Selected Lambda Numbers Values = ')
Lambda(M)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

BIBLIOGRAPHY

« Arizona State University Video Trace Library ». < http://trace.eas.asu.edu/yuv/ >. Accessed September 2012.

Bolin, M. R., and G. W. Meyer. 1999. « Visual difference metric for realistic image synthesis ». In *SPIE Human Vision and Electronic Imaging*. (San Jose, CA) Vol. 3644, p. 106-120.

Bovik, A. C. (553-596). 2009. *The essential guide to Image Processing*. USA: Academic Press.

Bystrom, M., I. Richardson and Y. Zhao. 2008. « Efficient mode selection for H.264 complexity reduction in a Bayesian framework ». *Signal Processing: Image Communication,* vol. 23, n° 2, p. 71-86.

Chandler, D. M., and S. S. Hemami. 2007. « VSNR: A wavelet-based visual signal-to-noise ratio for natural images ». *IEEE Transactions on Image Processing,* vol. 16, n° 9, p. 2284-2298.

Chen, L., and I. Garbacea. 2006. « Adaptive λ estimation in Lagrangian rate-distortion optimization for video coding ». In *Visual Communications and Image Processing (VCIP)*. (San Jose, USA, January) Vol. 6077, p. 60772.

Chen, Z., C. Du, J. Wang and Y. He. 2002. « PPFPS - a paraboloid prediction based fractional pixel search strategy for H.26L ». In *IEEE International Symposium on Circuits and Systems (ISCAS)*. (Scottsdale, Arizona, USA, May) Vol. 3, p. 9-12.

Chen, Z., and C. Guillemot. 2010. « Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, n° 6, p. 806-819.

Choi, I., J. Lee and B. Jeon. 2006. « Fast Coding Mode Selection With Rate-Distortion Optimization for MPEG-4 Part-10 AVC/H.264 ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 16, n° 12, p. 1557-1561.

Choi, J., and D. Park. 1994. « A stable feedback control of the buffer state using the controlled Lagrange multiplier method ». *IEEE Transactions on Image Processing,* vol. 3, n° 5, p. 546-558.

Chong, E. K. P., and S. H. Zak. 2011. *An Introduction to Optimization*, 3rd. Wiley-Interscience.

Rezazadeh, S., and S. Coulombe. US patent application 13/463,733. 2012. *Method and system for increasing robustness of visual quality metrics using spatial shifting*.

Daly, S. J. 1992. « The visible differences predictor: an algorithm for the assessment of image fidelity ». In *SPIE Human Vision, Visual Processing, and Digital Display*. (San Jose, CA, February) Vol. 1616, p. 2-15.

Damera-Venkata, N., T. D. Kite, W. S. Geisler, B. L. Evans and A. C. Bovik. 2000. « Image quality assessment based on a degradation model ». *IEEE Transactions on Image Processing,* vol. 9, n° 4, p. 636-650.

Daubechies, I. 1992. *Ten lectures on wavelets* (January), 1st. SIAM: Society for Industrial and Applied Mathematics.

Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264—ISO/IEC 14496-10 AVC). 2003. Geneva, Switzerland.

Gish, H., and J. Pierce. 1968. « Asymptotically efficient quantizing ». *IEEE Transactions on Information Theory,* vol. 14, n° 5, p. 676-683.

He, Z., and S. K. Mitra. 2002a. « A linear source model and a unified rate control algorithm for DCT video coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 12, n° 11, p. 970-982.

He, Z., and S. K. Mitra. 2002b. « Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 12, n° 10, p. 840-849.

Hrarti, M., H. Saadane, M-C. Larabi, A. Tamtaoui and D. Aboutajdine. 2010. « A macroblock-based perceptually adaptive bit allocation for H.264 rate control ». In *5th International Symposium on I/V Communications and Mobile Network (ISVC)* p. 1-4.

Huang, Y-H., T-S. Ou, P-Y. Su and H. H. Chen. 2010. « Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, n° 11, p. 1614-1624.

Huynh-Thu, Q., and M. Ghanbari. 2008. « Scope of validity of PSNR in image/video quality assessment ». *Electronics letters,* vol. 44, n° 13, p. 800-801.

Intel® 64 and IA32 architectures optimization reference manual. April 2012. Intel Corporation. < http://www.intel.com/content/dam/doc/manual/64-ia-32-architectures-optimization-manual.pdf >.

« Intel® Integrated Performance Primitives (Intel® IPP) ». < http://software.intel.com/en-us/intel-ipp/ >. Accessed January 2010.

ITU-T H.264 Telecommunication Standardization Sector of ITU. 2012. *Advanced video coding for generic audiovisual services*. Geneva.

Jeon, B., and J. Lee. 2003. *Fast mode decision for H.264*. Waikoloa, Hawaii, USA.

Joint Video Team (JVT) H.264/AVC Reference Software. < http://iphome.hhi.de/suehring/tml/ >. Accessed September 2012.

Kamaci, N., and Y. Altunbasak. 2004. « ρ-domain rate-distortion optimal rate control for DCT-based video coders ». In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. (Montreal, Canada, May) Vol. 3, p. 149-152.

Lai, Y-K., and C-C. J. Kuo. 2000. « A Haar wavelet approach to compressed image quality measurement ». *Journal of Visual Communication and Image Representation,* vol. 11, n° 1, p. 17-40.

Le Callet, P., and F. Autrusseau. 2005. « Subjective quality assessment IRCCyN/IVC database ». < http://www.irccyn.ec-nantes.fr/ivcdb >.

Li, W. , J-R. Ohm, M. van der Schaar, H. Jiang and S. Li. 2001. *MPEG-4 Video Verification Model version 18.0*. Pisa.

Li, X., N. Oertel, A. Hutter and A. Kaup. 2007. « Advanced Lagrange Multiplier Selection for Hybrid Video Coding ». In *IEEE International Conference on Multimedia and Expo*. (Beijing, China, July), p. 364-367.

Li, X., N. Oertel, A. Hutter and A. Kaup. 2009. « Laplace Distribution Based Lagrangian Rate Distortion Optimization for Hybrid Video Coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, n° 2, p. 193-205.

Li, Z. , F. Pan, K. P. Lim, G. Feng, X. Lin and S. Rahardja. 2003. *Adaptive basic unit layer rate control for JVT*. Pattaya II ,Thailand, 1-31 p.

Lim, K. P., G. Sullivan and T. Wiegand. 2005. *Text description of joint model reference encoding methods and decoding concealment methods*. Hong Kong.

Lin, G-X., and S-B. Zheng. 2008. « Perceptual importance analysis for H.264/AVC bit allocation ». *Journal of Zhejiang University Science A,* vol. 9, n° 2, p. 225-231.

Lin, Y-C., T. Fink and E. Bellers. 2007. « Fast Mode Decision for H.264 Based on Rate-Distortion Cost Estimation ». In *IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP). (Honolulu, Hawaii, USA, April) Vol. 1, p. 1137-1140.

Lubin, J. 1995. « A visual discrimination model for imaging system design and evaluation ». In *Vision models for target detection and recognition*. (Singapore) Vol. 2, p. 245-283.

Ma, S., W. Gao and D. Zhao. 2009. « Rate distortion cost modeling of skip mode and early skip mode selection for H. 264/MPEG-4 AVC ». In *SPIE Visual Communications and Image Processing*. (San Jose, CA, January) Vol. 7257. Society of Photo-Optical Instrumentation Engineers.

Mai, Z-Y., C-L. Yang, K-Z. Kuang and L-M. Po. 2006. « A Novel Motion Estimation Method Based on Structural Similarity for H.264 Inter Prediction ». In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Toulouse, France, May) Vol. 2, p. 913-916.

Mannos, J., and D. Sakrison. 1974. « The effects of a visual fidelity criterion of the encoding of images ». *IEEE Transactions on Information Theory,* vol. 20, n^{o} 4, p. 525-536.

Marpe, D., T. Wiegand and S. Gordon. 2005. « H.264/MPEG4-AVC fidelity range extensions: tools, profiles, performance, and application areas ». In *IEEE International Conference on Image Processing (ICIP)*. (Genoa, Italy, September) Vol. 1, p. 593-6.

Mitsa, T., and K. L. Varkur. 1993. « Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms ». In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5, p. 301-304.

Miyahara, M., K. Kotani and V. R. Algazi. 1998. « Objective picture quality scale (PQS) for image coding ». *IEEE Transactions on Communications,* vol. 46, n^{o} 9, p. 1215-1226.

OPAL Plug-In for Intel Integrated Performance Primitives. November 2012. < http://www.ippcodecs.org/index.html >.

Ortega, A., and K. Ramchandran. 1998. « Rate-distortion methods for image and video compression ». *IEEE Signal Processing Magazine,* vol. 15, n^{o} 6, p. 23-50.

Ou, T-S., Y-H. Huang and H. H. Chen. 2011. « SSIM-based perceptual rate control for video coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 21, n^{o} 5, p. 682-691.

Pan, F., X. Lin, R. Susanto, K. P. Lim, Z. G. Li, G. N. Feng, D. J. Wu and S. Wu. 2003. *Fast mode decision for intra prediction*. 6. Pattaya II, Thailand, 1-22 p.

Ponomarenko, N., V. Lukin, K. Egiazarian, J. Astola, M. Carli and F. Battisti. 2008. « Color image database for evaluation of image quality metrics ». In *IEEE 10th Workshop on Multimedia Signal Processing*. (Australia), p. 403-408.

Quan, D., and Y-S. Ho. 2010. « Categorization for fast intra prediction mode decision in H.264/AVC ». *IEEE Transactions on Consumer Electronics,* vol. 56, n° 2, p. 1049-1056.

Rezazadeh, S., and S. Coulombe. 2009. « A novel approach for computing and pooling structural similarity index in the discrete wavelet domain ». In *IEEE International Conference on Image Processing (ICIP)*. (November), p. 2209-2212.

Rezazadeh, S., and S. Coulombe. 2010. « Low-complexity computation of visual information fidelity in the discrete wavelet domain ». In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. (Dallas, TX, March), p. 2438-2441.

Rezazadeh, S., and S. Coulombe. 2011. « A novel discrete wavelet domain error-based image quality metric with enhanced perceptual performance ». In *18th International Conference on Signal Acquisition and Processing (ICSAP 2011)*. (Singapore, February) Vol. 1, p. 357-362.

Rezazadeh, S., and S. Coulombe. Granted US patent No. 8,326,046. 2012a. *Method and system for determining structural similarity between images*.

Rezazadeh, S., and S. Coulombe. 2012b. « A novel discrete wavelet domain error-based image quality metric with enhanced perceptual performance ». *International Journal of Computer and Electrical Engineering,* vol. 4, n° 2, p. 390-395.

Rezazadeh, S., and S. Coulombe. Granted US patent No. 8,515,181. 2013a. *Method and system for determining a quality measure for an image using a variable number of multi-level decompositions*.

Rezazadeh, S., and S. Coulombe. Granted US patent No. 8,515,182. 2013b. *Method and system for determining a quality measure for an image using multi-level decomposition of images*.

Rezazadeh, S., and S. Coulombe. Granted US patent continuation 13/692,950. 2013c. *Method and system for determining structural similarity between images*.

Rezazadeh, S., and S. Coulombe. 2013d. « A novel discrete wavelet transform framework for full reference image quality assessment ». *Signal, Image and Video Processing (Springer),* vol. 7, n° 3, p. 559-573.

Richardson, I. E. 2010. *The H.264 advanced video compression standard* (August), 2nd John Wiley & Sons Inc.

Rouse, D. M., and S. S. Hemami. 2008. « Understanding and simplifying the structural similarity metric ». In *IEEE International Conference on Image Processing*. (San Diego, CA, October), p. 1188-1191.

Safranek, R. J., and J. D. Johnston. 1989. « A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression ». In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)* (Glasgow, Scotland, May) Vol. 3, p. 1945-1948.

Sampat, M. P., Z. Wang, S. Gupta, A. C. Bovik and M. K. Markey. 2009. « Complex wavelet structural similarity: A new image similarity index ». *IEEE Transactions on Image Processing,* vol. 18, n° 11, p. 2385-2401.

Seshadrinathan, K., R. Soundararajan, A. C. Bovik and L. K. Cormack. 2010a. « Study of subjective and objective quality assessment of video ». *IEEE Transactions on Image Processing,* vol. 19, n° 6, p. 1427-1441.

Seshadrinathan, K., R. Soundararajan, A. C. Bovik and L. K. Cormack. 2010b. « A subjective study to evaluate video quality assessment algorithms ». In *SPIE Proceedings Human Vision and Electronic Imaging*. (San Jose, California January) Vol. 7527.

Sheikh, H. R., and A. C. Bovik. 2006. « Image information and visual quality ». *IEEE Transactions on Image Processing,* vol. 15, n° 2, p. 430-444.

Sheikh, H. R., A. C. Bovik and G. De Veciana. 2005. « An information fidelity criterion for image quality assessment using natural scene statistics ». *IEEE Transactions on Image Processing,* vol. 14, n° 12, p. 2117-2128.

Sheikh, H. R., M. F. Sabir and A. C. Bovik. 2006. « A statistical evaluation of recent full reference image quality assessment algorithms ». *IEEE Transactions on Image Processing,* vol. 15, n° 11, p. 3440-3451.

Sheikh, H. R., Z. Wang, L. K. Cormack and A. C. Bovik. « LIVE image quality assessment database Release 2 ». < http://live.ece.utexas.edu/research/quality/subjective.htm >.

Shen, L., Z. Liu, Z. Zhang and X. Shi. 2008. « Fast Inter Mode Decision Using Spatial Property of Motion Field ». *IEEE Transactions on Multimedia,* vol. 10, n° 6, p. 1208-1214.

Su, P-Y., C-K. Kao, T-Y. Huang and H. H. Chen. 2012. « Adopting Perceptual Quality Metrics in Video Encoders: Progress and Critiques ». In *IEEE International*

Conference on Multimedia and Expo Workshops (ICMEW). (Melbourne, Australia), p. 73-78.

Sullivan, G. , and G. Bjontegaard. 2001. *Recommended simulation common conditions for H.26L coding efficiency experiments on low-resolution progressive-scan source material*. 81. Santa Barbara, CA, USA.

Sullivan, G. J., and T. Wiegand. 1998. « Rate-distortion optimization for video compression ». *IEEE Signal Processing Magazine,* vol. 15, n° 6, p. 74-90.

Sun, C., H-J. Wang, T-H. Kim and H. Li. 2007. « Perceptually Adaptive Lagrange Multiplier for Rate-Distortion Optimization in H.264 ». In *Future Generation Communication and Networking (FGCN 2007)*. (Jeju-Island, Korea, December) Vol. 1, p. 459-463.

Sun, C., H-J. Wang and H. Li. 2008. « Macroblock-Level Rate-Distortion Optimization with Perceptual Adjustment for Video Coding ». In *Data Compression Conference (DCC)*. (Snowbird, Utah, March), p. 546-546.

Teo, P. C., and D. J. Heeger. 1994. « Perceptual image distortion ». In *IEEE International Conference on Image Processing*. (November) Vol. 2, p. 982-986.

Tourapis, A. M., H-Y. Cheong and P. Topiwala. 2005. *Fast ME in the JM reference software Status: Input Document to JVT Purpose: Proposal, Information*. Poznań, PL.

Tsukuba, T., I. Nagayoshi, T. Hanamura and H. Tominaga. 2005. « H.264 fast intra-prediction mode decision based on frequency characteristic ». In *13th European Signal Processing Conference (EUSIPCO 2005)*. (Antalya, Turkey, September) Vol. 2, p. 234-238.

Tu, Y-K., J-F. Yang and M-T. Sun. 2006. « Efficient rate-distortion estimation for H.264/AVC coders ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 16, n° 5, p. 600-611.

Video quality experts group. 2003. *Final report from the video quality experts group on the validation of objective models of video quality assessment*. < http://www.vqeg.org >.

Wang, H., S. Kwong and C-W. Kok. 2007. « An Efficient Mode Decision Algorithm for H.264/AVC Encoding Optimization ». *IEEE Transactions on Multimedia,* vol. 9, n° 4, p. 882-888.

Wang, S., S. Ma and W. Gao. 2010. « SSIM based perceptual distortion rate optimization coding ». In *SPIE Visual Communications and Image Processing (VCIP)*. (Huangshan, China, July) Vol. 7744.

Wang, S., A. Rehman, Z. Wang, S. Ma and W. Gao. 2011. « Rate-SSIM optimization for video coding ». In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Prague, Czech, May), p. 833-836.

Wang, S., A. Rehman, Z. Wang, S. Ma and W. Gao. 2012. « SSIM-Motivated Rate-Distortion Optimization for Video Coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, n° 4, p. 516-529.

Wang, Y., J. Ostermann and Y-Q. Zhang. 2002. *Video processing and communications*. New Jersey: Prentice-Hall.

Wang, Z. « The SSIM index for image quality assessment ». < https://ece.uwaterloo.ca/~z70wang/research/ssim/ >.

Wang, Z., and A. C. Bovik. 2006. *Modern image quality assessment*. Coll. « Synthesis Lectures on Image, Video, and Multimedia Processing ». United States: Morgan & Claypool, 1-156 p.

Wang, Z., and A. C. Bovik. 2009. « Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures ». *IEEE Signal Processing Magazine,* vol. 26, n° 1, p. 98-117.

Wang, Z., A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. 2004. « Image quality assessment: From error visibility to structural similarity ». *IEEE Transactions on Image Processing,* vol. 13, n° 4, p. 600-612.

Wang, Z., and Q. Li. 2011. « Information content weighting for perceptual image quality assessment ». *IEEE Transactions on Image Processing,* vol. 20, n° 5, p. 1185-1198.

Wang, Z., Q. Li and X. Shang. 2007. « Perceptual Image Coding Based on a Maximum of Minimal Structural Similarity Criterion ». In *IEEE International Conference on Image Processing (ICIP)*. (San Antonio, Texas, September) Vol. 2, p. 121-124.

Wang, Z., L. Lu and A. C. Bovik. 2004. « Video quality assessment based on structural distortion measurement ». *Signal Processing: Image Communication,* vol. 19, n° 2, p. 121-132.

Wang, Z., and X. Shang. 2006. « Spatial pooling strategies for perceptual image quality assessment ». In *IEEE International Conference on Image Processing*. (Atlanta, GA, October), p. 2945-2948.

Wang, Z., and E. P. Simoncelli. 2005. « Translation insensitive image similarity in complex wavelet domain ». In *IEEE International Conference on In Acoustics, Speech, and Signal Processing*. (March) Vol. 2, p. 573-576.

Wang, Z., E. P. Simoncelli and A. C. Bovik. 2003. « Multiscale structural similarity for image quality assessment ». In *IEEE Asilomar Conference on Signals, Systems and Computers*. (November) Vol. 2, p. 1398-1402.

Watson, A. B. 1993. « DCTune: A technique for visual optimization of DCT quantization matrices for individual images ». In *Society for Information Display Digest of Technical Papers*. Vol. 24, p. 946-949.

Wiegand, T. « H.264/AVC Video Coding Standard ». PDF document. < http://iphome.hhi.de/wiegand/assets/pdfs/DIC_H264_07.pdf >. Accessed 2012, July 26.

Wiegand, T., and B. Girod. 2001. « Lagrange multiplier selection in hybrid video coder control ». In *IEEE International Conference on Image Processing*. (Thessaloniki , Greece, October) Vol. 3, p. 542-545.

Wiegand, T., M. Lightstone, D. Mukherjee, T. G. Campbell and S. K. Mitra. 1996. « Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 6, n° 2, p. 182-190.

Wiegand, T., H. Schwarz, A. Joch, F. Kossentini and G. J. Sullivan. 2003a. « Rate-constrained coder control and comparison of video coding standards ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 13, n° 7, p. 688-703.

Wiegand, T., and G. Sullivan. 2003. *Recommendation and final draft international standard of joint video specification (ITU-T Rec. H. 264| ISO/IEC 14496-10 AVC)*. Pattaya, Thailand.

Wiegand, T., G. J. Sullivan, G. Bjontegaard and A. Luthra. 2003b. « Overview of the H.264/AVC video coding standard ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 13, n° 7, p. 560-576.

Wien, M. 2003. « Variable block-size transforms for H.264/AVC ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 13, n° 7, p. 604-613.

x264: a Free H.264/MPEG-4 AVC Software Library and Application. 2012. < http://www.videolan.org/developers/x264.html >. Accessed November 2012.

Xin, J., A. Vetro and H. Sun. 2004. « Efficient macroblock coding-mode decision for H. 264/AVC video coding ». In *Picture Coding Symposium (PCS)*. (San Francisco, USA, December), p. 15-17. Citeseer.

Yang, C-L., W-R. Gao and L-M. Po. 2008. « Discrete wavelet transform-based structural similarity for image quality assessment ». In *IEEE International Conference on Image Processing*. (San Diego, CA, October), p. 377-380.

Yang, C-L., R-K. Leung, L-M. Po and Z-Y. Mai. 2009. « An SSIM-optimal H.264/AVC inter frame encoder ». In *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*. (Shanghai, November) Vol. 4, p. 291-295.

Yang, C. L., H. X. Wang and L. M. Po. 2007. « A novel fast motion estimation algorithm based on SSIM for H.264 video coding ». In *Advances in Multimedia Information Processing (PCM 2007)*. (Berlin). Vol. 4810, p. 168-176.

Yasakethu, S. L. P., W. A. C. Fernando, S. Adedoyin and A. Kondoz. 2008. « A rate control technique for offline H.264/AVC video coding using subjective quality of video ». *IEEE Transactions on Consumer Electronics,* vol. 54, n^{o} 3, p. 1465-1472.

Yeo, C., H. L. Tan and Y. H. Tan. 2012. « On rate distortion optimization using SSIM ». In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Kyoto, Japan, March), p. 833-836.

Yu, H., F. Pan, Z. Lin and Y. Sun. 2005. « A perceptual bit allocation scheme for H.264 ». In *IEEE International Conference on Multimedia and Expo (ICME)*. (Amsterdam, The Netherlands, July).

Yuan, W., S. Lin, Y. Zhang, W. Yuan and H. Luo. 2006. « Optimum bit allocation and rate control for H.264/AVC ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 16, n^{o} 6, p. 705-715.

Zhang, J., X. Yi, N. Ling and W. Shang. 2006. *Context adaptive Lagrange multiplier (CALM) for motion estimation in JM-improvement*. Klagenfurt, Austria.

Zhang, J., X. Yi, N. Ling and W. Shang. 2007. « Chroma Coding Efficiency Improvement with Context Adaptive Lagrange Multiplier (CALM) ». In *IEEE International Symposium on Circuits and Systems (ISCAS)*. (New Orleans, May), p. 293-296.

Zhang, J., X. Yi, N. Ling and W. Shang. 2010. « Context Adaptive Lagrange Multiplier (CALM) for Rate-Distortion Optimal Motion Estimation in Video Coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, n^{o} 6, p. 820-828.